

<sup>1</sup>

# EXPERIMENTOLOGY

<sup>2</sup>

AN OPEN SCIENCE APPROACH TO EXPERIMENTAL PSYCHOLOGY METHODS

<sup>3</sup>

MICHAEL C. FRANK      MIKA BRAGINSKY      JULIE CACHIA

<sup>4</sup>

NICHOLAS A. COLES      TOM E. HARDWICKE

<sup>5</sup>

ROBERT D. HAWKINS      MAYA B. MATHUR

<sup>6</sup>

RONDELLE WILLIAMS

<sup>7</sup>

2024-10-08

# <sup>8</sup> CONTENTS

<sup>9</sup>	PREFACE	5
<sup>10</sup>	I FOUNDATIONS	21
<sup>11</sup>	1 EXPERIMENTS	22
<sup>12</sup>	2 THEORIES	43
<sup>13</sup>	3 REPLICATION	71
<sup>14</sup>	4 ETHICS	120
<sup>15</sup>	II STATISTICS	147
<sup>16</sup>	5 ESTIMATION	153
<sup>17</sup>	6 INFERENCE	180
<sup>18</sup>	7 MODELS	228

19	III PLANNING	271
20	8 MEASUREMENT	275
21	9 DESIGN	326
22	10 SAMPLING	373
23	IV EXECUTION	411
24	11 PREREGISTRATION	417
25	12 DATA COLLECTION	448
26	13 PROJECT MANAGEMENT	495
27	V REPORTING	531
28	14 WRITING	537
29	15 VISUALIZATION	570
30	16 META-ANALYSIS	636
31	17 CONCLUSION	672
32	APPENDICES	678
33	A INSTRUCTOR'S GUIDE	680
34	B GIT AND GITHUB (ONLINE ONLY)	704
35	C R MARKDOWN AND QUARTO (ONLINE ONLY)	705
36	D TIDYVERSE (ONLINE ONLY)	706

E	GGPLOT (ONLINE ONLY)	707
---	----------------------	-----

## <sup>38</sup> PREFACE

<sup>39</sup> As scientists and practitioners, we often want to create generalizable,  
<sup>40</sup> causal theories of human behavior. As it turns out, experiments—in  
<sup>41</sup> which we use random assignment to measure a causal effect—are an  
<sup>42</sup> unreasonably effective tool to help with this task. But how should we  
<sup>43</sup> go about doing good experiments?

<sup>44</sup> This book provides an introduction to the workflow of the experimen-  
<sup>45</sup> tal researcher working in psychology or the behavioral sciences more  
<sup>46</sup> broadly. The organization of the book is sequential, from the plan-  
<sup>47</sup> ning stages of the research process through design, data gathering, anal-  
<sup>48</sup> ysis, and reporting. We introduce these concepts via narrative examples  
<sup>49</sup> from a range of subdisciplines, including cognitive, developmental, and  
<sup>50</sup> social psychology. Throughout, we also illustrate the pitfalls that led to  
<sup>51</sup> the “replication crisis” in psychology.

52 To help researchers avoid these pitfalls, we advocate for an open science–  
53 based approach in which transparency is integral to the entire experi-  
54 mental workflow. We provide readers with guidance for preregistra-  
55 tion, project management, data sharing, and reproducible report writ-  
56 ing.

57 *The story of this book*

58 Experimental Methods (Psych 251) is the foundational course for in-  
59 coming graduate students in the Stanford psychology department. The  
60 course goal is to orient students to the nuts and bolts of doing behav-  
61 ioral experiments, including how to plan and design a solid experiment  
62 and how to avoid common pitfalls regarding design, measurement, and  
63 sampling.

64 Almost all student coursework both before and in graduate school deals  
65 with the content of their research, including theories and results in their  
66 areas of focus. In contrast, our course is sometimes the only one that  
67 deals with the *process* of research, from big questions about why we do  
68 experiments and what it means to make a causal inference all the way  
69 to the tiny details of project organization, such as what to name your  
70 directories and how to make sure you don't lose your data in a computer  
71 crash.

<sup>72</sup> This observation leads to our book’s title. “Experimentology” is the  
<sup>73</sup> set of practices, findings, and approaches that enable the construction of  
<sup>74</sup> robust, precise, and generalizable experiments.

<sup>75</sup> The centerpiece of the Experimental Methods course is a replication  
<sup>76</sup> project, reflecting a teaching model first described in Frank and Saxe  
<sup>77</sup> (2012) and later expanded on in Hawkins et al. (2018). Each student  
<sup>78</sup> chooses a published experiment in the literature and collects new data  
<sup>79</sup> on a preregistered version of the same experimental paradigm, compar-  
<sup>80</sup> ing their result to the original publication. Over the course of the quar-  
<sup>81</sup> ter, we walk through how to set up a replication experiment, how to  
<sup>82</sup> preregister confirmatory analyses, and how to write a reproducible re-  
<sup>83</sup> port on the findings. The project teaches concepts like reliability and  
<sup>84</sup> validity, which allow students to analyze choices that the original exper-  
<sup>85</sup> imenters made—often choices that could have been made differently in  
<sup>86</sup> hindsight!

<sup>87</sup> At the end of the course, we reap the harvest of these projects. The  
<sup>88</sup> project presentations are a wonderful demonstration of both how much  
<sup>89</sup> the students can accomplish in a quarter and how tricky it can be to  
<sup>90</sup> reproduce (redo calculations in the original data) and replicate (recover  
<sup>91</sup> similar results in new data) the published literature. Often our repli-  
<sup>92</sup> cation success rate for the course hovers just above 50%, an outcome

<sup>93</sup> that can be disturbing or distressing for students who assume that the  
<sup>94</sup> published literature reports the absolute truth.

<sup>95</sup> This book is an attempt to distill some of the lessons of the course (and  
<sup>96</sup> students' course projects) into a textbook. We'll tell the story of the  
<sup>97</sup> major shifts in psychology that have come about in the last ten years,  
<sup>98</sup> including both the "replication crisis" and the positive methodological  
<sup>99</sup> reforms that have resulted from it. Using this story as motivation, we  
<sup>100</sup> will highlight the importance of transparency during all aspects of the  
<sup>101</sup> experimental process from planning to dissemination of materials, data,  
<sup>102</sup> and code.

### <sup>103</sup> *What this book is and isn't about*

<sup>104</sup> This book is about psychology experiments. These will typically be  
<sup>105</sup> short studies conducted online or in a single visit to a lab, often—though  
<sup>106</sup> certainly not always—with a convenience sample. When we say "exper-  
<sup>107</sup> iments" here, we mean **randomized experiments** where some aspect of  
<sup>108</sup> the participants' experience is **manipulated** by the experimenter and  
<sup>109</sup> then some outcome variable is **measured**.<sup>1</sup>

<sup>110</sup> The central thesis of the book is that:

<sup>1</sup> We use **bold** to indicate the introduc-  
tion of new technical terms.

111 Experiments are intended to make maximally unbiased,  
112 generalizable, and precise estimates of specific causal  
113 effects.

114 We'll explore the implications of this thesis for a host of topics, includ-  
115 ing causal inference, experimental design, measurement, sampling, pre-  
116 registration, data analysis, and many others.

117 Because our focus is on experiments, we won't be talking much about  
118 observational designs, survey methods, or qualitative research; these  
119 are important tools and are appropriate for a whole host of questions,  
120 but they aren't our focus here. We also won't go into depth about  
121 the many fascinating methodological and statistical issues brought up  
122 by single-participant case studies, longitudinal research, field studies, or  
123 other methodological variants. Many of the concerns we raise are still  
124 important for these types of studies, but some of our advice won't trans-  
125 fer to these less common designs.

126 Even for students who are working on nonexperimental research, we  
127 expect that a substantial part of the book content will still be useful,  
128 including chapters on replication (chapter 3), ethics (chapter 4), statis-  
129 tics (chapters 5, 6, 7), sampling (chapter 10), project management (chap-  
130 ter 13), and reporting (chapters 14, 15, 16).

<sup>131</sup> In our writing, we presuppose that readers have some background in  
<sup>132</sup> psychology, at least at an introductory level. In addition, although we  
<sup>133</sup> introduce a number of statistical topics, readers might find these sections  
<sup>134</sup> more accessible with an undergraduate statistics course under their belt.  
<sup>135</sup> Finally, our examples are written in the R statistical programming lan-  
<sup>136</sup> guage, and for chapters on statistics and visualization especially (chap-  
<sup>137</sup> ters 5, 6, 7, 15, 16), some familiarity with R will be helpful for under-  
<sup>138</sup> standing the code. None of these prerequisites are necessary to read  
<sup>139</sup> the book, but we offer them so that readers can calibrate their expecta-  
<sup>140</sup> tions.

### <sup>141</sup> *How to use this book*

<sup>142</sup> The book is organized into five main parts, mirroring the timeline of  
<sup>143</sup> an experiment: (1) “Foundations,” (2) “Statistics,” (3) “Planning,” (4)  
<sup>144</sup> “Execution,” and (5) “Reporting.” We hope that this organization  
<sup>145</sup> makes it well suited for teaching or for use as a reference book.<sup>2</sup>

<sup>146</sup> The book is designed for a course for graduate students or advanced un-  
<sup>147</sup> dergraduates, but the material is also suitable for self-study by anyone  
<sup>148</sup> interested in experimental methods, whether in academic psychology  
<sup>149</sup> or any other context—in or out of academia—in which behavioral ex-  
<sup>150</sup> perimentation is relevant. We also hope that some readers will come to

<sup>2</sup> If you are an instructor who is planning to adopt the book for a course, you might be interested in our resources for instructors, including sample course schedules, in appendix A.

<sup>151</sup> particular chapters of the book because of an interest in specific topics  
<sup>152</sup> like measurement (chapter 8) or sampling (chapter 10) and will be able  
<sup>153</sup> to use those chapters as stand-alone references. And finally, for those  
<sup>154</sup> interested in the “replication crisis” and subsequent reforms, chapters 3,  
<sup>155</sup> 11, and 13 will be especially interesting.

<sup>156</sup> Ultimately, we want to give you what you need to plan and execute  
<sup>157</sup> your own study! Instead of enumerating different approaches, we try  
<sup>158</sup> to provide a single coherent—and often quite opinionated—perspective,  
<sup>159</sup> using marginal notes and references to give pointers to more advanced  
<sup>160</sup> materials or alternative approaches. Throughout, we offer:

<sup>161</sup> – **CASE STUDIES** that illustrate the central concepts of a chapter  
<sup>162</sup> – **ACCIDENT REPORTS** describing examples where poor research prac-  
<sup>163</sup> tices led to issues in the literature  
<sup>164</sup> – **DEPTH BOXES** providing simulations, linkages to advanced tech-  
<sup>165</sup> niques, or more nuanced discussion

<sup>166</sup> While case studies are often integral to the chapters, the other boxes  
<sup>167</sup> can typically be skipped without issue.

<sup>168</sup> *Themes*

<sup>169</sup> We highlight four major cross-cutting themes for the book: TRANSPARENCY, MEASUREMENT PRECISION, BIAS REDUCTION, and GENERALIZABILITY.<sup>3</sup>

<sup>3</sup> Themes are noted using SMALL CAPS.

<sup>172</sup> — TRANSPARENCY: For experiments to be reproducible, other researchers need to be able to determine exactly what you did.

<sup>174</sup> Thus, every stage of the research process should be guided by a primary concern for transparency. For example, preregistration creates transparency into the researcher's evolving expectations and thought processes; releasing open materials and analysis scripts creates transparency into the details of the procedure.

<sup>179</sup> — MEASUREMENT PRECISION: We want researchers to start planning an experiment by thinking: "What causal effect do I want to measure?" and to make planning, sampling, design, and analytic choices that maximize the precision of this measurement. A downstream consequence of this mindset is that we move away from a focus on dichotomized inferences about statistical significance and toward analytic and meta-analytic models that focus on continuous effect sizes and confidence intervals.

<sup>187</sup> — BIAS REDUCTION: While precision refers to random error in a measurement, measurements also have systematic sources of error that

189 bias them away from the true quantity. In our samples, analyses,  
190 and experimental designs, and in the literature, we need to think  
191 carefully about sources of bias in the quantity being estimated.

- 192 – GENERALIZABILITY: Complex behaviors are rarely universal across  
193 all settings and populations, and any given experiment can only  
194 hope to cover a small slice of the possible conditions where a be-  
195 havior of interest takes place. Psychologists must therefore con-  
196 sider the generalizability of their findings at every stage of the  
197 process, from stimulus selection and sampling procedures to ana-  
198 lytic methods and reporting.

199 Throughout the book, we will return to these four themes again and  
200 again as we discuss how the decisions made by the experimenter at ev-  
201 ery stage of design, data gathering, and analysis bear on the inferences  
202 that can be made about the results. The introduction of each chapter  
203 highlights connections to specific themes.

204 *The software toolkit for this book*

205 We introduce and advocate for an approach to reproducible study plan-  
206 ning, analysis, and writing. This approach depends on an ecosystem of  
207 open-source software tools, which we introduce in the book’s appen-  
208 dices:<sup>4</sup>

<sup>4</sup> These appendices are available online at <https://experimentology.io> but not in the print version of the book, since their content is best viewed in the web format.

- 209 – the R statistical programming language and the RStudio<sup>5</sup> inte-  
210 grated development environment  
211 – version control using git and GitHub<sup>6</sup> for allowing collaboration  
212 on text documents like code, prose, and data, storing and integrat-  
213 ing contributions over time (appendix B)  
214 – the RMarkdown and Quarto tools for creating reproducible re-  
215 ports that can be rendered to a variety of formats (appendix C)  
216 – the tidyverse family of R packages, which extend the basic func-  
217 tionality of R with simple tools for data wrangling, analysis, and  
218 visualization (appendix D)  
219 – the ggplot2 plotting package, which makes it easy to create flexi-  
220 ble data visualizations for both confirmatory and exploratory data  
221 analyses (appendix E)

222 Where appropriate, we provide **code boxes** that show the specific R  
223 code used to create our examples.

224 *Onward!*

225 Thanks for joining us for *Experimentology*! Whether you are casually  
226 browsing, doing readings for a course, or using the book as a reference in  
227 your own experimental work, we hope you find it useful. Throughout,  
228 we have tried to practice what we preach in terms of reproducibility,

<sup>5</sup> <https://posit.co/download/rstudio-desktop>

<sup>6</sup> <https://github.com>

<sup>229</sup> and so the full source code for the book is available at <https://github.com/langcog/experimentology>. We encourage you to browse, com-<sup>230</sup>  
<sup>231</sup> ment, and log issues or suggestions.<sup>7</sup>

<sup>7</sup> The best way to give us specific feedback is to create an issue on our github page at <https://github.com/langcog/experimentology/issues>.

## <sup>232</sup> *Acknowledgments*

<sup>233</sup> Thanks first and foremost to the many generations of students and TAs  
<sup>234</sup> in Stanford Psych 251, who have collectively influenced the content of  
<sup>235</sup> this book through their questions and interests.

<sup>236</sup> Thanks to the staff at the MIT Press, especially Philip Laughlin and Amy  
<sup>237</sup> Brand, for embracing a vision of a completely open web textbook that is  
<sup>238</sup> also reviewed and published through a traditional press. We appreciate  
<sup>239</sup> your support and flexibility.

<sup>240</sup> We adapt the Contributor Roles (CRediT) Taxonomy<sup>8</sup> to describe our  
<sup>241</sup> contributions to this manuscript, and we recommend that you do so in  
<sup>242</sup> your work as well.

<sup>8</sup> Learn more at <https://credit.niso.org>.

### <sup>243</sup> – Preface

<sup>244</sup> – Primary writer: Michael C. Frank  
<sup>245</sup> – Editor: Tom E. Hardwicke

### <sup>246</sup> – Chapter 1

247 – Primary writer: Michael C. Frank

248 – Cowriter: Nicholas Coles

249 – Editor: Tom E. Hardwicke

250 – Chapter 2

251 – Primary writer: Michael C. Frank

252 – Editors: Nicholas Coles and Tom E. Hardwicke

253 – Chapter 3

254 – Primary writer: Michael C. Frank

255 – Editors: Maya B. Mathur, Tom E. Hardwicke, and Nicholas

256 Coles

257 – Chapter 4

258 – Primary writer: Rondeline Williams

259 – Cowriter: Michael C. Frank

260 – Editors: Tom E. Hardwicke and Julie Cachia

261 – Chapter 5

262 – Cowriters: Maya B. Mathur, Nicholas Coles, and Michael

263 C. Frank

264 – Editors: Julie Cachia and Tom E. Hardwicke

265 – Chapter 6

266 – Primary writer: Michael C. Frank

267 – Cowriter: Maya B. Mathur

268 – Editors: Julie Cachia and Tom E. Hardwicke

269 – Chapter 7

270 – Cowriters: Maya B. Mathur and Michael C. Frank

271 – Editors: Tom E. Hardwicke and Robert D. Hawkins

272 – Chapter 8

273 – Primary writer: Michael C. Frank

274 – Editors: Robert D. Hawkins, Tom E. Hardwicke, and Ron-

275 deline Williams

276 – Chapter 9

277 – Primary writer: Michael C. Frank

278 – Editors: Nicholas Coles and Tom E. Hardwicke

279 – Chapter 10

280 – Primary writer: Michael C. Frank

281 – Editors: Julie Cachia, Tom E. Hardwicke, and Maya B.

282 Mathur

283 – Chapter 11

284 – Primary writer: Tom E. Hardwicke

285 – Editor: Michael C. Frank

286 – Chapter 12

287 – Cowriters: Rondeline Williams and Michael C. Frank

288 – Editor: Tom E. Hardwicke

289 – Chapter 13

290 – Primary writer: Michael C. Frank

291 – Editor: Tom E. Hardwicke

292 – Chapter 15

293 – Primary writer: Robert D. Hawkins

294 – Editors: Michael C. Frank, Tom E. Hardwicke, and Mika

295 Braginsky

296 – Chapter 14

297 – Primary writer: Tom E. Hardwicke

298 – Editor: Michael C. Frank

299 – Chapter 16

300 – Cowriters: Nicholas Coles and Maya B. Mathur

301 – Editors: Michael C. Frank and Tom E. Hardwicke

302 – Conclusion

303 – Primary writer: Michael C. Frank

304 – Editor: Tom E. Hardwicke

305 – Appendix A

306 – Primary writer: Julie Cachia

307 – Editor: Michael C. Frank

308 – Appendix B

309 – Primary writer: Julie Cachia

310 – Editor: Michael C. Frank

311 – Appendix C

312 – Primary writer: Michael C. Frank

313 – Editor: Julie Cachia

314 – Appendix D

315 – Primary writer: Michael C. Frank

316 – Editors: Julie Cachia and Mika Braginsky

317 – Appendix E

318 – Primary writer: Michael C. Frank

319 – Editors: Julie Cachia and Mika Braginsky

320 – Technical infrastructure

321 – Developers: Mika Braginsky and Natalie Braginsky

<sup>322</sup> *References*

Frank, Michael C., and Rebecca Saxe. 2012. “Teaching Replication.” *Perspectives on Psychological Science* 7 (6): 600–604. <https://doi.org/10.1177/1745691612460686>.

Hawkins, Robert D., Eric N. Smith, Carolyn Au, Juan Miguel Arias, Rhia Cat-  
<sup>323</sup> apano, Eric Hermann, Martin Keil, et al. 2018. “Improving the Replica-  
bility of Psychological Science Through Pedagogy.” *Advances in Methods  
and Practices in Psychological Science* 1 (1): 7–18. <https://doi.org/10.1177/2515245917740427>.

<sup>324</sup>

325

|

326

## FOUNDATIONS

# 1 EXPERIMENTS

327

## APPLE LEARNING GOALS

- Define what an experiment is
- Contrast observational and experimental studies using causal graphs
- Understand the role of randomization in experiments
- Consider constraints on the generalizability of experiments

328

329 Welcome to *Experimentology*! This is a book all about the art of running  
330 experiments in psychology. Throughout, we will be guided by a simple  
331 idea:

332 The purpose of experiments is to estimate the magnitude  
333 of causal effects.<sup>1</sup>

334 Starting from our core idea, we'll provide advice about how to navi-  
335 gate things like experimental design, measurement, sampling, and more.

<sup>1</sup> Perhaps you're already saying, "That's not what I thought experiments were for! I thought they were for testing hypotheses." Bear with us and we hope we'll convince you that our definition is a bit more general, and that testing a hypothesis is one thing you can do with a measurement.

336 Our decisions about each of these will determine how precise our es-  
337 timate is, and whether it is subject to bias. But before we get to those  
338 topics, let's start by thinking about *why* we might do an experiment, a  
339 topic that will intersect with our key themes of BIAS REDUCTION and  
340 GENERALIZABILITY.

### 341 1.1 *Observational studies don't reveal causality*

342 If you're reading this book, there's probably something about psychol-  
343 ogy you want to understand. How is language learned? How is it that  
344 we experience emotions like happiness and sadness? Why do humans  
345 sometimes work together and other times destroy one another? When  
346 psychologists study these centuries-old questions, they often transform  
347 them into questions about **causality**.<sup>2</sup>

#### 348 1.1.1 *Describing causal relationships*

349 Consider the age-old question: Does money make people happy? This  
350 question is—at its heart—a question about what interventions on the  
351 world we can make. Can I get more money and make myself happier?  
352 Can I *cause* happiness with money?

<sup>2</sup> Defining causality is one of the trickiest and oldest problems in philosophy, and we won't attempt to solve it here! But from a psychological perspective, we're fond of Lewis's (1973) "counterfactual" analysis of causality. On this view, we can understand the claim that *money causes happiness* by considering a scenario where if people *hadn't* been given more money, they *wouldn't* have experienced an increase in happiness.

353 How could we test our hypothesized effect of money on happiness?  
 354 Intuitively, many people think of running an **observational study**. We  
 355 might survey people about how much money they make and how happy  
 356 they are. The result of this study would be a pair of measurements for  
 357 each participant: [money, happiness].

358 Now, imagine your observational study found that money and happiness  
 359 were related—statistically **correlated** with each another: people with  
 360 more money tended to be happier. Can we conclude that money causes  
 361 happiness? Not necessarily. The presence of a correlation does not  
 362 mean that there is a causal relationship!

363 Let's get a bit more precise about our causal hypothesis. To illustrate  
 364 causal relationships, we can use a tool called **directed acyclic graphs**  
 365 (DAGs, Pearl 1998). Figure 1.1 shows an example of a DAG for money  
 366 and happiness. The arrow represents our idea about the potential causal  
 367 link between two variables: money and happiness.<sup>3</sup> The direction of  
 368 the arrow tells us which way we hypothesize that the causal relation-  
 369 ship goes.

370 The correlation between money and happiness we saw in our observa-  
 371 tional study is consistent with the causal model in figure 1.1; however,  
 372 it is also consistent with several alternative causal models, which we will  
 373 illustrate with DAGs below.

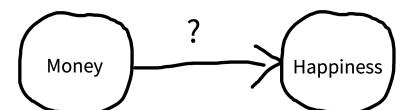


Figure 1.1  
 The hypothesized causal effect of money on happiness.

<sup>3</sup> In this chapter, we're going to use the term “variables” without discussing why we study some variables and not others. In the next chapter, we'll introduce the term “construct,” which indicates a psychological entity that we want to theorize about.

<sup>374</sup> 1.1.2 *The problems of directionality and confounding*

<sup>375</sup> Figure 1.2 uses DAGs to illustrate several causal models that are consistent with the observed correlation between money and happiness. DAG 1 represents our hypothesized relationship—money causes people to be happy. But DAG 2 shows an effect in completely the opposite direction! In this DAG, being happy causes people to make more money.

<sup>380</sup> Even more puzzling, there could be a correlation but no causal relationship between money and happiness in either direction. Instead, a third variable—often referred to as a **confound**—may be causing increases in both money and happiness. For example, maybe having more friends causes people to both be happier and make more money (DAG 3). In this scenario, happiness and money would be correlated even though one does not cause the other.

<sup>387</sup> A confound (or several) may entirely explain the relationship between two variables (as in DAG 3), but it can also just *partly* explain the relationship. For example, it could be that money does increase happiness, but the causal effect is rather small, and only accounts for a small portion of the observed correlation between them, with the friendship confound (and perhaps others) accounting for the remainder.

<sup>393</sup> In this case, because of the confounds, we say that the observed corre-

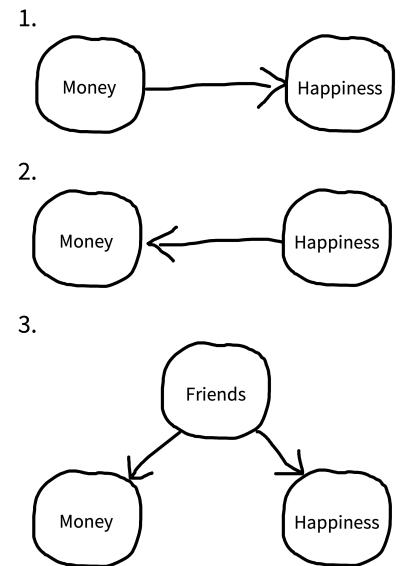


Figure 1.2  
Three reasons why money and happiness can be correlated.

394 lation between money and happiness is a **biased** estimate of the causal  
395 effect of money on happiness. The amount of bias introduced by the  
396 confounds can vary in different scenarios—it may only be small or it  
397 may be so strong that we conclude that there's a causal relationship be-  
398 tween two variables when there isn't one at all.

399 The state of affairs summarized in figure 1.2 is why we say “correlation  
400 doesn't imply causation.” A correlation between two variables *is consist-*  
401 *tent with* a causal relationship between them, but it's also consistent with  
402 other relationships as well.<sup>4</sup>

403 You can still learn about causal relationships from observational stud-  
404 ies, but you have to take a more sophisticated approach. You can't just  
405 measure correlations and leap to causal conclusions. The “causal rev-  
406 olution” in the social sciences has been fueled by the development of  
407 statistical methods for reasoning about causal relationships from obser-  
408 vational datasets.<sup>5</sup> As interesting as these methods are, however, they  
409 are only applicable in certain specific circumstances. In contrast, the  
410 experimental method *always* works to reduce bias due to confounding  
411 (though of course there are certain experiments that we can't do for  
412 ethical or practical reasons).

<sup>4</sup> People sometimes ask whether *causa-*  
*tion implies correlation* (the opposite di-  
rection). The short answer is “also no.”  
A causal relationship between two vari-  
ables often means that they will be corre-  
lated in the data, but not always. For ex-  
ample, imagine you measured the speed  
of a car and the pressure on the gas  
pedal / accelerator. In general, pres-  
sure and speed will be correlated, con-  
sistent with the causal relationship be-  
tween the two. But now imagine you  
only measured these two variables when  
someone was driving the car up a hill—  
now the speed would be constant, but  
the pressure might be increasing, reflect-  
ing the driver's attempts to keep their  
speed up. So there would be no corre-  
lation between the two variables in that  
dataset, despite the continued causal re-  
lationship.

## 413 1.2 Experiments help us answer causal questions

414 Imagine that you (a) created an exact replica of our world, (b) gave  
415 \$1,000 to everybody in the replica world, and then (c) found a few years  
416 later that everyone in the replica world was happier than their matched  
417 self in the original world. This experiment would provide strong evi-  
418 dence that money makes people happier. Let's think through why.

419 Consider a particular person—if they are happier in the replica that in  
420 the original world, what could explain that difference? Since we have  
421 replicated the world exactly but made only one change—money—then  
422 that change is the only factor that could explain the difference in hap-  
423 piness. We can say that we **held all variables constant** except for money,  
424 which we **manipulated** experimentally, observing its effect on some  
425 **measure**—happiness. This idea—holding all variables constant except  
426 for the specific experimental manipulation—is the basic logic that un-  
427 derpins the experimental method (as articulated by Mill 1843).<sup>6</sup> Let's  
428 think back to our observational study of money and happiness. One  
429 big causal inference problem was the presence of “third-variable” con-  
430 founds like having more friends. More friends could cause you to have  
431 more money and also cause you to be happier. The idea of an ex-  
432 periment is to hold everything else constant—including the number of  
433 friends that people have—so we can measure the effect of money on

<sup>5</sup> In fact, DAGs are one of the key tools that social scientists use to reason about causal relationships. DAGs guide the creation of statistical models to estimate particular causal effects from observational data. We won't talk about these methods here, but if you're interested, check out the suggested readings at the end of this chapter.

<sup>6</sup> Another way to reason about why we can infer causality here follows the counterfactual logic we described in an earlier footnote. If the definition of causality is counterfactual (“What would have happened if the cause had been different?”), then this experiment fulfills that definition. In our impossible experiment, we can literally *see* the counterfactual: if the person had \$1,000 more, here's how much happier they would be!

434 happiness. By holding the number of friends constant, we would be  
 435 severing the causal links between friends and both money and happi-  
 436 ness. This move is graphically conveyed in figure 1.3, where we “snip  
 437 away” the friend confound.

### 438 1.2.1 *We can't hold people constant*

439 This all sounds great in theory, you might be thinking, but we can't  
 440 actually create replica worlds where everything is held constant, so how  
 441 do we run experiments in the real world? If we were talking about  
 442 experiments on baking cakes, it's easy to see how we could hold all of  
 443 the ingredients constant and just vary one thing, like baking temperature.  
 444 Doing so would allow us to conduct an experimental test of the effect of  
 445 baking temperature. But how can we “hold something constant” when  
 446 we're talking about people? People aren't cakes. No two people are  
 447 alike, and, as every parent with multiple children knows, even if you try  
 448 to “hold the ingredients constant,” they don't come out the same!

449 If we take two people and give one money, we're comparing two *differ-*  
 450 *ent* people, not two instances of the same person with everything held  
 451 constant. It wouldn't work to *make* the first person have more or fewer  
 452 friends so they match the second person—that's not holding anything

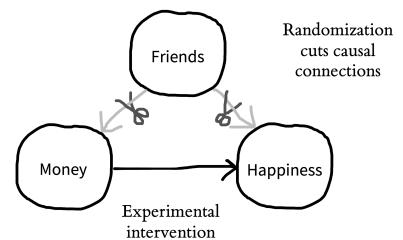


Figure 1.3

In principle, experiments allow us to “snip away” the friend confound by holding it constant (though in practice, it can be tough to figure out how to hold something constant when you are talking about people as your unit of study).

453 constant; instead it's another (big, difficult, and potentially unethical)  
 454 intervention that might itself cause lots of effects on happiness.

455 You may be wondering: Why don't we just ask people how many  
 456 friends they have and use this information to split them into equal  
 457 groups? You could do that, but this kind of strategy only allows you  
 458 to control for the confounds you know of. For example, you may  
 459 split people equally based on their number of friends but not their  
 460 educational attainment. If educational attainment also impacts both  
 461 money and happiness, you still have a confound. You may then try  
 462 to split people by both their number of friends and their education.  
 463 But perhaps there's another confound you've missed: sleep quality!  
 464 Similarly, it also doesn't work to select people who have the same  
 465 number of friends—that only holds the friends variable constant and  
 466 not everything *else* that's different between the two people. So what  
 467 do we do instead?<sup>7</sup>

#### 468 1.2.2 Randomization saves the day

469 The answer is **randomization**. If you randomly split a large roomful  
 470 of people into two groups, the groups will, on average, have a similar  
 471 number of friends. Similarly, if you randomly pick who in your experi-  
 472 ment gets to receive money, you will find that the money and no-money

<sup>7</sup> Many researchers who have seen regression models used in the social sciences assume that “controlling for lots of stuff” is a good way to improve causal inference. Not so! In fact, inappropriately controlling for a variable in the absence of a clear causal justification can actually make your effect estimate *more* biased (Wysocki, Lawson, and Rheumtulla 2022).

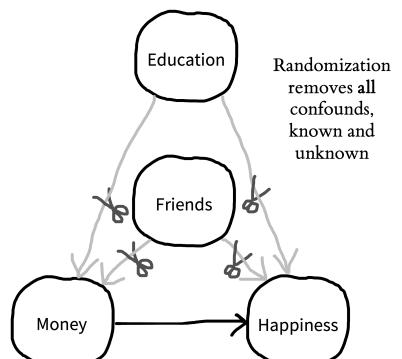


Figure 1.4  
 If you randomly split a large group of people into groups, the groups will, on average, be equal in every way.

473 groups, on average, have a similar number of friends. In other words,  
474 through randomization, the confounding role of friends is controlled.  
475 But the most important thing is that it's not *just* the role of friends that's  
476 controlled; educational attainment, sleep quality, and all the other con-  
477 founders are controlled as well. If you randomly split a large group of  
478 people into groups, the groups will, on average, be equal in every way  
479 (figure 1.4).

480 So, here's our simple experimental design: we randomly assign some  
481 people to a money group and some people to a no-money control group  
482 (we sometimes call these groups **conditions**). Then we measure the hap-  
483 piness of people in both groups. The basic logic of randomization is that,  
484 if money causes happiness, we should see more happiness—on average—  
485 in the money group.<sup>8</sup>

486 Randomization is a powerful tool, but there is a caveat: it doesn't work  
487 every time. *On average*, randomization will ensure that your money and  
488 no-money groups will be equal with respect to confounds like number  
489 of friends, educational attainment, and sleep quality. But just as you can  
490 flip a coin and sometimes get heads nine out of ten times, sometimes  
491 you use randomization and still get more highly educated people in one  
492 condition than the other. When you randomize, you guarantee that,  
493 on average, all confounds are controlled. Hence, there is no systematic

<sup>8</sup> You may already be protesting that this experiment could be done better. Maybe we could measure happiness before and after randomization, to increase precision. Maybe we need to give a small amount of money to participants in the control condition to make sure that participants in both conditions interact with an experimenter, and hence, that the conditions are as similar as possible. We agree! These are important parts of experimental design, and we'll touch on them in subsequent chapters.

<sup>494</sup> bias in your estimate from these confounds. But there will still be some  
<sup>495</sup> noise from random variation.

<sup>496</sup> In sum, randomization is a remarkably simple and effective way of hold-  
<sup>497</sup> ing everything constant besides a manipulated variable. In doing so,  
<sup>498</sup> randomization allows experimental psychologists to make unbiased es-  
<sup>499</sup> timates of causal relationships. Importantly, randomization works both  
<sup>500</sup> when you do have control of every aspect of the experiment—like when  
<sup>501</sup> you are baking a cake—and even when you don’t—like when you are  
<sup>502</sup> doing experiments with people.<sup>9</sup>

#### DEPTH

### *Unhappy randomization?*

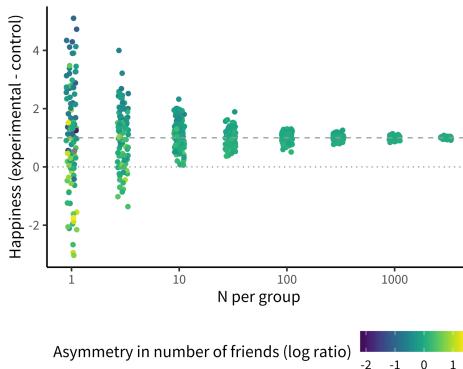
As we’ve been discussing, random assignment removes confounding by ensuring that—on average—groups are equivalent with respect to all of their characteristics. Equivalence for any *particular* random assignment is more likely the larger your sample is, however. Any individual experiment may be affected by **unhappy randomization**, when a particular confound is unbalanced between groups by chance.

Unhappy randomization is much more common in small experiments than larger ones. To see why, we use a technique called **simulation**. In simulations, we invent data randomly following a set of assumptions: we make up a group of participants and generate their characteristics and

<sup>9</sup> There’s an important caveat to this discussion: you don’t always have to randomize *people*. You can use an experimental design called a **within-participants** design, in which the same people are in multiple conditions. This type of design has a different set of unknown confounds, this time centering around *time*. So, to get around them, you have to randomize the order in which your manipulation is delivered. This randomization works very well for some kinds of manipulations but not so well for others. We’ll talk more about these kinds of designs in chapter 9.

their condition assignments. By varying the assumptions we use, we can investigate how particular choices might change the structure of the data.

To look at unhappy randomization, we created many simulated versions of our money-happiness experiment, in which an experimental group receives money and the control group receives none, and then happiness is measured for both groups. We assume that each participant has a set number of friends, and that the more friends they have, the happier they are. So, when we randomly assign them to experimental and control groups, we run the risk of unhappy randomization—sometimes one group will have substantially more friends than the other.



**Figure 1.5**  
Simulated data from our money-happiness experiment. Each dot represents the measured happiness effect (vertical position) for an experiment with a set number of participants in each group (horizontal position). The dot color shows how uneven friendship is between the groups. The dashed line shows the true effect.

Figure 1.5 shows the results of this simulation. Each dot is an experiment, representing one estimate of the happiness effect (how much happiness is gained for the amount of money given to the experimental group). For very small experiments (e.g., with one or three participants per group), dots are very far from the dashed line showing the true effect—meaning

these estimates are extremely noisy! And the reason is unhappy randomization. The upper and lower points are those in which one group had far more friends than the other.

There are three things to notice about this simulation. First, the noise overall goes down as the sample sizes get bigger: larger experiments yield estimates closer to the true effect. Second, the unhappy randomization decreases dramatically as well with larger samples. Although individuals still differ just as much in large experiments, the *group* average number of friends is virtually identical for each condition in the largest groups.

Finally, although the small experiments are individually very noisy, the *average effect* across all of the small experiments is still very close to the true effect. This last point illustrates what we mean when we say that randomized experiments remove confounds. Even though friendship is still an important factor determining happiness in our simulation, the average effect across experiments is correct and each individual estimate is unbiased.

505

### 506 1.3 Generalizability

507 When we are asking questions about psychology, it's important to think  
508 about who we are trying to study. Do we want to know if money in-  
509 creases happiness in *all people*? In people who live in materialistic so-  
510 cieties? In people whose basic needs are not being met? We call the

511 group we are trying to study our **population of interest** and the people  
512 who actually participate in our experiment our **sample**. The process of  
513 **sampling** is then what we do to recruit people into our experiment.

514 Sometimes researchers sample from one population but make a claim  
515 about another (usually broader) population. For example, they may run  
516 their experiment with a particular sample of US college students but  
517 then generalize to all people (their intended population of interest). The  
518 mismatch of sample and population is not always a problem, but quite  
519 often causal relationships are different for different populations.

520 Unfortunately, psychologists pervasively assume that research on US  
521 and European samples generalizes to the rest of the world, and it of-  
522 ten does not. To highlight this issue, Henrich, Heine, and Norenzayan  
523 (2010) coined the acronym WEIRD. This catchy name describes the  
524 oddness of making generalizations about all of humanity from exper-  
525 iments on a sample that is quite unusual because it is Western, Edu-  
526 cated, Industrialized, Rich, and Democratic. Henrich and colleagues  
527 argue that seemingly “fundamental” psychological functions like visual  
528 perception, spatial cognition, and social reasoning all differ pervasively  
529 across populations—hence, any generalization from an effect estimated  
530 with a WEIRD subpopulation may be unwarranted.

531 In the early 2000s, researchers found that gratitude interventions—like

532 writing a brief essay about something nice that somebody did for you—  
533 increased happiness in studies conducted in Western countries. Based on  
534 these findings, some psychologists believed that gratitude interventions  
535 could increase happiness in all people. But it seems they were wrong.  
536 A few years later, Layous et al. (2013) ran a gratitude experiment in  
537 two locations: the United States and South Korea. Surprisingly, the  
538 gratitude intervention decreased happiness in the South Korean sample.  
539 The researchers attributed this negative effect to feelings of indebted-  
540 ness that people in South Korea more prominently experienced when  
541 reflecting on gratitude. In this example, we would say that the findings  
542 obtained with the US sample may not **generalize** to people in South  
543 Korea.

544 Issues of generalizability extend to all aspects of an experiment, not just  
545 its sample. For example, even if our hypothetical cash intervention ex-  
546 periment resulted in gains in happiness, we might not be warranted in  
547 generalizing to different ways of providing money. Perhaps there was  
548 something special about the amount of money we gave or the way we  
549 provided it that led to the effect we observed. Without testing multi-  
550 ple different intervention types, we can't make a broad claim. As we'll  
551 see in chapter 7 and chapter 9, this issue has consequences for both our  
552 statistical analyses and our experimental designs (Yarkoni 2020).

553 Questions of generalizability are pervasive, but the first step is to simply  
 554 acknowledge and reason about them. Perhaps all papers should have a  
 555 Constraints on Generality statement, where researchers discuss whether  
 556 they expect their findings to generalize across different samples, exper-  
 557 imental stimuli, procedures, and historical and temporal features (Si-  
 558 mons, Shoda, and Lindsay 2017). This kind of statement would at least  
 559 remind researchers to be humble: experiments are a powerful tool for  
 560 understanding how the world works, but there are limits to what any  
 561 individual experiment can teach us.

562 *1.4 Anatomy of a randomized experiment*

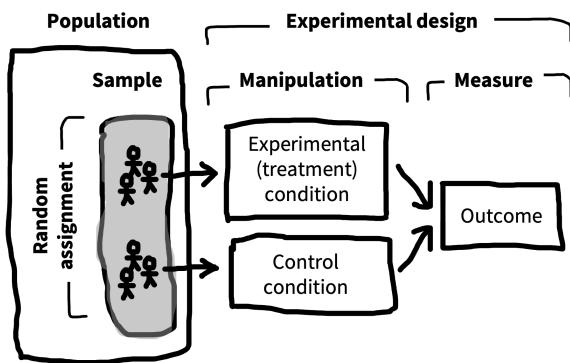


Figure 1.6  
 The anatomy of a randomized experiment.

563 Now is a good time for us to go back and consolidate the anatomy of  
 564 an experiment, since this anatomy is used throughout the book. Fig-  
 565 ure 1.6 shows a simple two-group experiment like our possible money-  
 566 happiness intervention. A sample is taken from a larger population, and  
 567 then participants in the sample are randomly assigned to one of two

568 conditions (the manipulation)—either the experimental condition, in  
569 which money is provided, or the control condition, in which none is  
570 given. Then an outcome measure—happiness—is recorded for each par-  
571 ticipant.

572 We'll have a lot more to say about all of these components in subsequent  
573 chapters. We'll discuss measures in chapter 8, because good measure-  
574 ment is the foundation of a good experiment. Then in chapter 9, we'll  
575 discuss the different kinds of experimental designs that are possible and  
576 their pros and cons. Finally, we'll cover the process of sampling in chap-  
577 ter 10.

### ⚠ ACCIDENT REPORT

#### *An experiment with very unclear causal inferences*

The Stanford Prison Experiment is one of the most famous studies in the history of psychology. Participants were randomly assigned to play the role of “guards” and “prisoners” in a simulation of prison life inside the Stanford Department of Psychology building ([Zimbardo 1972](#)). Designed to run for two weeks, the simulation had to be ended after six days due to the cruelty of the participants acting as guards, who apparently engaged in a variety of dehumanizing behaviors toward the simulated prisoners. This result is widely featured in introductory psychology textbooks and is typically interpreted as showing the power of situational factors: in the right context, even undergraduate students at Stanford could quickly be

convinced to act out the kind of inhumane behaviors found in the worst prisons in the world (Griggs 2014).

In the years since the study was initially reported, a variety of information has surfaced that makes the causal interpretation of its situational manipulation much less clear (Le Texier 2019). Guards were informed of the objectives of the experiment and given instructions on how to achieve these objectives. The experimenters themselves suggested some harsh punishments whose later use was given as evidence for the emergence of dehumanizing behaviors. Further, both guards and prisoners were coached extensively by the experimenter throughout the study. Some participants have reported that their responses during the study were exaggerated or fabricated (Blum 2018). All of these issues substantially undermine the idea that the assignment of participants' roles (the ostensible experimental manipulation) was the sole cause of the observed behaviors.

The conduct of the study was also unethical. In addition to the question of whether such a study—with all of its risks to the participants—would be ethical at all, a number of features of the study clearly violate the guidelines that we'll learn about in chapter 4. Participants were prevented from exiting the study voluntarily. The guards were deceived into believing that they were research assistants, rather than participants in the study. And to top it off, the study was reported inaccurately, with reports emphasizing the organic emergence of behaviors, the immersive nature of the simulation, and the extensive documentation of the experiment. In fact, the participants were instructed extensively, the simulation was

repeatedly interrupted by mundane details of the research environment, and relatively little of the experiment was captured on video and analyzed.

The Prison Experiment is a fascinating and problematic episode in the history of psychology, but it provides very little causal evidence about the human mind.

580

### 581 1.5 *Chapter summary: Experiments*

582 In this chapter, we defined an experiment as a combination of a ma-  
583 nipulation and a measure. When combined with randomization, ex-  
584 periments allow us to make strong causal inferences, even when we are  
585 studying people (who are hard to hold constant). Nonetheless, there are  
586 limits to the power of experiments: there are always constraints on the  
587 sample, experimental stimuli, and procedure that limit how broadly we  
588 can generalize.



#### DISCUSSION QUESTIONS

1. Imagine that you run a survey and find that people who spend more time playing violent video games tend to be more aggressive (i.e., that there is a positive correlation between violent video games and aggression). Following figure 1.2, list three reasons why these variables may be correlated.
2. Suppose you wanted to run an experiment testing whether playing

589

violent video games causes increases in aggression. What would be your manipulation, and what would be your measure? How would you deal with potential confounding by variables like age?

3. Consider an experiment designed to test people's food preferences.

The experimenter randomly assigns thirty US preschoolers to be served either asparagus or chicken tenders and then asks them how much they enjoyed their meal. Overall, children enjoyed the meat more; the experimenter writes a paper claiming that humans prefer meat over vegetables. List some constraints on the generalizability of this study. In light of these constraints, is this study (or some modification) worth doing at all?

4. Consider the Milgram study, another classic psychology study (and our case study in chapter 4). Does this study meet our definition of an experiment?

590

## READINGS

- A basic introduction to causal inference from a social science perspective: Huntington-Klein, N. (2022). *The Effect: An Introduction to Research Design and Causality*. Chapman & Hall. Available free online at <https://theeffectbook.net>.
- A slightly more advanced treatment, focusing primarily on econometrics: Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale Press.

591

Available free online at <https://mixtape.scunning.com>.

592

## 593 References

- Blum, Ben. 2018. “The Lifespan of a Lie.” Medium. <https://medium.com/s/trustissues/the-lifespan-of-a-lie-d869212b1f62>.
- Griggs, Richard A. 2014. “Coverage of the Stanford Prison Experiment in Introductory Psychology Textbooks.” *Teaching of Psychology* 41 (3): 195–203.
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?” *Behavioral and Brain Sciences* 33 (2-3): 61–83.
- Layous, Kristin, Hyunjung Lee, Incheol Choi, and Sonja Lyubomirsky. 2013. “Culture Matters When Designing a Successful Happiness-Increasing Activity: A Comparison of the United States and South Korea.” *Journal of Cross-Cultural Psychology* 44 (8): 1294–1303.
- Le Texier, Thibault. 2019. “Debunking the Stanford Prison Experiment.” *American Psychologist* 74 (7): 823.
- Lewis, David. 1973. *Counterfactuals*. John Wiley & Sons.
- Mill, John Stuart. 1843. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. John W. Parker.
- Pearl, Judea. 1998. “Graphical Models for Probabilistic and Causal Reasoning.” *Quantified Representation of Uncertainty and Imprecision*, 367–389.
- Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay. 2017. “Constraints on Generality (COG): A Proposed Addition to All Empirical Papers.” *Per-*

594

- spectives on Psychological Science* 12 (6): 1123–1128.
- Wysocki, Anna C., Katherine M. Lawson, and Mijke Rhemtulla. 2022. “Statistical Control Requires Causal Justification.” *Advances in Methods and Practices in Psychological Science* 5 (2): 25152459221095823.
- Yarkoni, Tal. 2020. “The Generalizability Crisis.” *Behavioral and Brain Sciences* 45:1–37.
- 595 Zimbardo, Philip G. 1972. “Pathology of Imprisonment.” *Society* 9 (6): 4–8.

## 2 THEORIES

596



### LEARNING GOALS

- Define theories and their components
- Contrast different philosophical views on scientific theories
- Analyze features of an experiment that can lead to strong tests of theory
- Discuss the role of formalization in theory development

597

598 When you do an experiment, sometimes you just want to see what hap-  
599 pens, like a kid knocking down a tower made of blocks. And sometimes  
600 you want to know the answer to a specific applied question, like “Will  
601 giving a midterm vs weekly quizzes lead students in a class to perform  
602 better on the final?” But more often, our goal is to create **theories** that  
603 help us explain and predict new observations.

604 What is a theory? We’ll argue here that we should think of psycholog-  
605 ical theories as sets of proposed relationships among **constructs**, which  
606 are variables that we think play causal roles in determining behavior. In

607 this conception of theories, the role of causality is central: theories are  
608 guesses about the causal structure of the mind and about the causal re-  
609 lationships between the mind and the world. This definition doesn't  
610 include everything that gets called a "theory" in psychology. We de-  
611 scribe the continuum between theories and **frameworks**—broad sets of  
612 ideas that guide research but don't make specific contact with particular  
613 empirical observations.

614 We begin this chapter by talking about the specific enterprise of con-  
615 structing psychological theories. We'll then discuss how theories make  
616 contact with data, reviewing a bit of the philosophy of science, and give  
617 some guidance on how to construct experiments that test theories. We  
618 end by discussing the relationship between theories and quantitative  
619 models. This material touches on several of our book themes, includ-  
620 ing **GENERALIZABILITY** of theories and the need for **MEASUREMENT PRE-**  
621 **CISION** to make strong tests of theory.

## 622 *2.1 What is a psychological theory?*

623 The definition we just gave for a psychological theory is that it is a pro-  
624 posed set of causal relationships among constructs that helps us explain  
625 behavior. Let's look at the ingredients of a theory: the constructs and

626 the relationships between them. Then we can ask about how this defi-  
627 nition relates to other things that get called “theories” in psychology.

628 *2.1.1 Psychological constructs*

629 Constructs are the psychological variables that we want our theory to  
630 describe, like “money” and “happiness” in the example from last chap-  
631 ter. At first glance, it might seem odd that we need a specific name for  
632 these variables. But in probing the relationship between money and  
633 happiness, we will have to figure out a way to measure happiness. Let’s  
634 say we just ask people to answer the question “How happy are you?” by  
635 giving ratings on a 1 (miserable) to 10 (elated) scale.

636 Now say someone in the study reports they are an 8 on this scale. Is this  
637 *really* how happy they are? What if they weren’t concentrating very  
638 hard on the rating, or if they thought the researcher wanted them to  
639 be happy? What if they act much less happy in their interactions with  
640 family and friends?

641 We resolve this dilemma by saying that the self-report ratings we collect  
642 are only a **measure** of a **latent** construct, happiness. The construct is  
643 latent because we can never see it directly, but we think it has a causal  
644 influence on the measure: happier people should, on average, provide  
645 higher ratings. But many other factors can lead to noise or bias in the

646 measurement, so we shouldn't mistake those ratings as actually *being* the  
647 construct.

648 The particular question "How happy are you?" is one way of going  
649 from the general construct to a specific measure. The general process  
650 of going from construct to a specific instantiation that can be measured  
651 or manipulated is called **operationalization**. Happiness can be opera-  
652 tionalized by self-report, but it can also be operationalized many other  
653 ways—for example, through a measure like the use of positive language  
654 in a personal essay, or by ratings by friends, family, or a clinician. These  
655 decisions about how to operationalize a construct with a particular mea-  
656 sure are tricky and consequential, and we discuss them extensively in  
657 chapter 8. Each different operationalization might be appropriate for a  
658 specific study, yet it would require some justification and argument to  
659 connect each one to the others.

660 Proposing a particular construct is a very important part of making a  
661 theory. For example, a researcher might worry that self-reported hap-  
662 piness is very different than someone's well-being as observed by the  
663 people around them and assert that happiness is not a single construct  
664 but rather a group of distinct constructs. This researcher would then  
665 be surprised to know that self-reports of happiness relate very highly  
666 to others' perceptions of a person's well-being (Sandvik, Diener, and

667 Seidlitz 1993).<sup>1</sup>

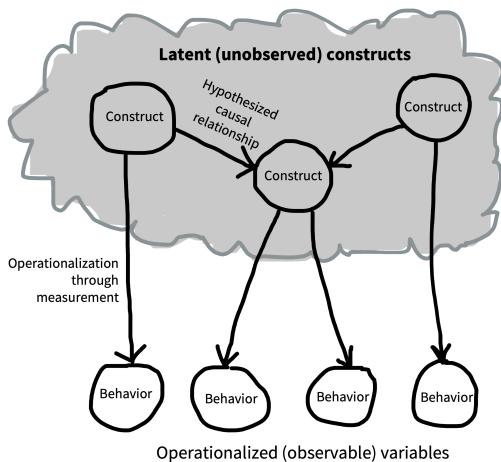
668 Even external, apparently nonpsychological variables like money don't  
669 have direct effects on people but rather operate through psychological  
670 constructs. People studying money seriously as a part of psychological  
671 theories think about perceptions of money in different ways depending  
672 on the context. For example, researchers have written about the im-  
673 portance of how much money you have on hand based on when in the  
674 month your paycheck arrives (Ellwood-Lowe, Foushee, and Srinivasan  
675 2022) but have also considered perceptions of long-term accumulation  
676 of wealth as a way of conceptualizing people's understanding of the dif-  
677 ferent resources available to White and Black families in the United  
678 States (Kraus et al. 2019).

679 Finally, a construct can be operationalized through a manipulation: in  
680 our money-happiness example, we operationalized "more money" in  
681 our theory with a gift of a specific amount of cash. We hope you see  
682 through these examples that operationalization is a huge part of the craft  
683 of being a psychology researcher—taking a set of abstract constructs that  
684 you're interested in and turning them into a specific experiment with a  
685 manipulation and a measure that tests your causal theory. We'll have a  
686 lot more to say about how this is done in chapter 9.

<sup>1</sup> Sometimes positing the construct *is* the key part of a theory. *g* (general intelligence) is the classic psychological example of a single-construct theory. The idea behind *g* theory is that the best measure of general intelligence is the shared variance between a wide variety of different tests. The decision to theorize about and measure a single unified construct for intelligence—rather than say, many different separate kinds of intelligence—is itself a controversial move.

687 2.1.2 *The relationships between constructs*

688 Constructs gain their meaning in part via their own definitions and op-  
689 erationalizations, but also in part through their causal relationships to  
690 other constructs. Figure 2.1 shows a schematic of what this kind of the-  
691 ory might look like—as you can see, it looks a lot like the DAGs that we  
692 introduced in the last chapter! That’s no accident. The arrows here also  
693 describe hypothesized causal links.<sup>2</sup>



694 This web of constructs and assumptions is what Cronbach and Meehl  
695 (1955) referred to as a “nomological network”—a set of proposals about  
696 how different entities are connected to one another. The tricky part  
697 is that the key constructs are never observed directly. They are in peo-  
698 ple’s heads.<sup>3</sup> So researchers only get to probe them by measuring them  
699 through specific operationalizations.

700 One poetic way of thinking about this idea is that the theoretical system

2 Sometimes these kind of diagrams are used in the context of a statistical method called structural equation modeling (SEM), where circles represent constructs and lines represent their relationships with one another. Confusingly, structural equation models are also used by many researchers to describe psychological theories. The important point for now is that they are one particular statistical formalism, not a general tool for theory building—the points we are trying to make here are more general.

3 We’re not saying these should correspond to specific brain structures. They could, but most likely they won’t. The idea that psychological constructs are not the same as any particular brain state (and especially not any particular brain region) is called “multiple realizability” by philosophers, who mostly agree that psychological states can’t be reduced to brain states, as much as philosophers agree on anything (Block and Fodor 1972).

701 of constructs “floats … above the plane of observation and is anchored  
702 to it by the rules of **measurement**.” ([Hempel 1952](#)). So, even if your  
703 theory posits that two constructs (say, money and happiness) are directly  
704 related, the best you can do is manipulate one operationalization and  
705 measure another operationalization. If this manipulation doesn’t pro-  
706 duce any effect, it’s possible that you are wrong and money does not  
707 cause happiness—but it is also possible that your operationalizations are  
708 poor.

709 Here’s a slightly different way of thinking about a theory. A theory  
710 provides a **compression** of potentially complex data into much a smaller  
711 set of general factors. If you have a long sequence of numbers, say [2 4 8  
712 16 32 64 128 256 …], then the expression  $2^n$  serves as a compression of  
713 this sequence—it’s a short expression that tells you what numbers are in  
714 vs out of the sequence. In the same way, a theory can compress a large  
715 set of observations (maybe data from many experiments) into a small set  
716 of relationships between constructs. Now, if your data are noisy, say [2.2  
717 3.9 8.1 16.1 31.7 …], then the theory will not be a perfect representation  
718 of the data. But it will still be useful.

719 In particular, having a theory allows you to **explain** observed data and  
720 **predict** new data. Both of these are good things for a theory to do.  
721 For example, if it turned out that the money causes happiness theory

722 was true, we could use it to explain observations such as greater levels  
723 of happiness among wealthy people. We could also make predictions  
724 about the effects of policies like giving out a universal basic income on  
725 overall happiness.<sup>4</sup> Explanation is an important feature of good theories,  
726 but it's also easy to trick yourself by using a vague theory to explain a  
727 finding post hoc (after the fact). Thus, the best test of a theory is typi-  
728 cally a new prediction, as we discuss below.

729 One final note: Causal diagrams are a very useful formalism, but they  
730 leave the generalizability of the causal relationships implicit. For ex-  
731 ample, will more money result in more happiness for everyone, or just  
732 for people at particular ages or in particular cultural contexts? “Who  
733 does this theory apply to?” is an important question to ask about any  
734 proposed causal framework.

<sup>4</sup> The relationship between money and happiness is actually much more complicated than what we're assuming here. For example, Killingsworth, Kahneman, and Mellers (2023) describes a collaboration between two sets of researchers that had different viewpoints on the connection between money and happiness.

### 735 2.1.3 *Specific theories vs general frameworks*

736 You may be thinking, “Psychology is full of theories but they don’t look  
737 that much like the ones you’re talking about!” Very few of the theories  
738 that bear this label in psychology describe causal relationships linking  
739 clearly defined and operationalized constructs. You also don’t see that  
740 many DAGs, though these are getting (slightly) more common lately  
741 (Rohrer 2018).

742 Here's an example of something that gets called a theory yet doesn't  
 743 share the components described above. Bronfenbrenner's (1992) eco-  
 744 logical systems theory (EST) is pictured in figure 2.2. The key thesis  
 745 of this theory is that children's development occurs in a set of nested  
 746 contexts that each affect one another and in turn affect the child. This  
 747 theory has been immensely influential. Yet, if it's read as a causal the-  
 748 ory, it's almost meaningless: nearly everything connects to everything  
 749 in both directions, and the constructs are not operationalized—it's very  
 750 hard to figure out what kind of predictions it makes!

751 Ecological systems theory is not really a theory in the sense that we are  
 752 advocating for in this chapter—and the same goes for many other very  
 753 interesting ideas in psychology. It's not a set of causal relationships be-  
 754 tween constructs that allow specific predictions about future observa-  
 755 tions. Ecological systems theory is instead a broad set of ideas about  
 756 what sorts of theories are more likely to explain specific phenomena.  
 757 For example, it helps remind us that a child's behavior is likely to be  
 758 influenced by a huge range of factors, such that any individual theory  
 759 cannot just focus on an individual factor and hope to provide a full ex-  
 760 planation. In this sense, EST is a **framework**: it guides and inspires spe-  
 761 cific theories—in the sense we've discussed here, namely a set of causal  
 762 relationships between constructs—without being a theory itself.

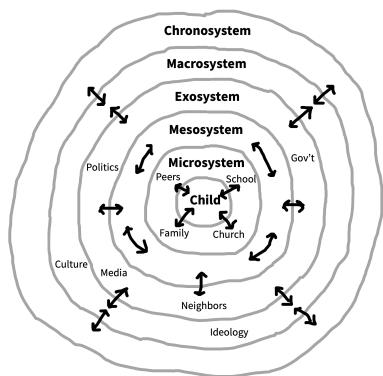


Figure 2.2

The diagram often used to represent Bronfenbrenner's ecological systems theory. Note that circles no longer denote discrete constructs; arrows can be interpreted as causal relationships, but all constructs are assumed to be fully connected.

763 Frameworks like EST are often incredibly important. They can also  
764 make a big difference to practice. For example, EST supports a model  
765 in social work in which children's needs are considered not only as the  
766 expression of specific internal developmental issues but also as stemming  
767 from a set of overlapping contextual factors (Ungar 2002). Concretely, a  
768 therapist might be more likely to examine family, peer, and school envi-  
769 ronments when analyzing a child's situation through the lens of EST.

770 There's a continuum between precisely specified theories and broad  
771 frameworks. Some theories propose interconnected constructs but  
772 don't specify the relationships between them, or don't specify how  
773 those constructs should be operationalized. So when you read a paper  
774 that says it proposes a "theory," it's a good idea to ask whether it  
775 describes specific relations between operationalized constructs. If it  
776 doesn't, it may be more of a framework than a theory.

### ↗ ACCIDENT REPORT

#### *The cost of a bad theory*

Theory development isn't just about knowledge for knowledge's sake—it has implications for the technologies and policies built off the theories.

One case study comes from Edward Clarke's infamous theory regarding the deleterious effects of education for women. Clarke posited that (1) cognitive and reproductive processes relied on the same fixed pool of en-

ergy, (2) relative to men, women's reproductive processes required more energy, and that (3) expending too much energy on cognitive tasks like education depleted women of the energy needed to maintain a healthy reproductive system. Based on case studies, Clarke suggested that education was causing women to become ill, experience fertility issues, and birth weaker children. He thus concluded that "boys must study and work in a boy's way, and girls in a girl's way." (Clarke 1884, p. 19).

Clarke's work is a chilling example of the implication of a poorly developed theory. In this scenario, Clarke had neither instruments that allowed him to measure his constructs nor experiments to measure the causal connections between them. Instead, he merely highlighted case studies that were consistent with his idea (while simultaneously dismissing cases that were inconsistent). His ideas eventually lost favor—especially as they were subjected to more rigorous tests. But Clarke's arguments were used to attempt to dissuade women from pursuing higher education and hindered educational policy reform.

778

## *779 2.2 How do we test theories?*

780 Our view of psychological theories is that they describe a set of rela-  
781 tionships between different constructs. How can we test theories and  
782 decide which one is best? We'll first describe **falsificationism**, a histori-  
783 cal viewpoint on this issue that has been very influential in the past and

<sup>784</sup> that connects to ideas about statistical inference presented in chapter 6.

<sup>785</sup> We'll then turn to a more modern viewpoint, **holism**, that recognizes

<sup>786</sup> the interconnections between theory and measurement.

### <sup>787</sup> 2.2.1 Falsificationism

<sup>788</sup> One historical view that resonates with many scientists is the philoso-

<sup>789</sup> pher Karl Popper's **falsificationism**. In particular, there is a simplistic

<sup>790</sup> version of falsificationism that is often repeated by working scientists,

<sup>791</sup> even though it's much less nuanced than what Popper actually said! On

<sup>792</sup> this view, a scientific theory is a set of hypotheses about the world that

<sup>793</sup> instantiate claims like the connection between money and happiness.<sup>5</sup>

<sup>794</sup> What makes a statement a *scientific* hypothesis is that it can be disproved

<sup>795</sup> (i.e., it is **falsifiable**) by an observation that contradicts it. For example,

<sup>796</sup> observing a lottery winner who immediately becomes depressed would

<sup>797</sup> falsify the hypothesis that receiving money makes you happier.

<sup>798</sup> For the simplistic falsificationist, theories are never **confirmed**. The

<sup>799</sup> hypotheses that form parts of theories are universal statements. You

<sup>800</sup> can never prove them right; you can only fail to find falsifying evidence.

<sup>801</sup> Seeing hundreds of people get happier when they received money

<sup>802</sup> would not prove that the money-happiness hypothesis was universally

<sup>803</sup> true. There could always be a counter-example around the corner.

<sup>5</sup> Earlier we treated the claim that money caused happiness as a theory. It is one! It's just a very simple theory that has only one hypothesized connection in it.

804 This theory doesn't really describe how scientists work. For example,  
805 scientists like to say that their evidence "supports" or "confirms" their  
806 theory, and falsificationism rejects this kind of talk. A falsificationist  
807 says that confirmation is an illusion; that the theory is simply surviving  
808 to be tested another day. This strict falsificationist perspective is unpalat-  
809 able to many scientists. After all, if we observe that hundreds of people  
810 get happier when they receive money, it seems like this should at least  
811 slightly increase our confidence that money causes happiness!<sup>6</sup>

### 812 2.2.2 *A holistic viewpoint on theory testing*

813 The key issue that leads us to reject strict falsificationism is the obser-  
814 vation that no individual hypothesis (a part of a theory) can be falsi-  
815 fied independently. Instead, a large series of what are called **auxiliary**  
816 **assumptions** (or auxilliary hypotheses) are usually necessary to link an  
817 observation to a theory (Lakatos 1976). For example, if giving some  
818 individual person money didn't change their happiness, we wouldn't  
819 immediately throw out our theory that money causes happiness. In-  
820 stead, the fault might be in any one of our auxiliary assumptions, like  
821 our measurement of happiness, or our choice of how much money to  
822 give or when to give it. The idea that individual parts of a theory can't  
823 be falsified independently is sometimes called **holism**.

<sup>6</sup> An alternative perspective comes from the Bayesian tradition that we'll learn more about in chapters 5 and 6. In a nutshell, Bayesians propose that our subjective belief in a particular hypothesis can be captured by a probability, and that our scientific reasoning can then be described by a process of normative probabilistic reasoning (Strevens 2006). The Bayesian scientist distributes probability across a wide range of alternative hypotheses; observations that are more consistent with a hypothesis increase the hypothesis's probability (Sprenger and Hartmann 2019).

824 One consequence of holism is that the relationship between data and  
825 theory isn't always straightforward. An unexpected observation may  
826 not cause us to give up on a main hypothesis in our theory—but it will  
827 often cause us to question our auxiliary assumptions instead (e.g., how  
828 we operationalize our constructs). Thus, before abandoning our theory  
829 of money causing happiness, we might want to try several happiness  
830 questionnaires.

831 The broader idea of holism is supported by historical and sociological  
832 studies of how science progresses, especially in the work of Kuhn (1962).  
833 Examining historical evidence, Kuhn found that scientific revolutions  
834 didn't seem to be caused by the falsification of a theoretical statement via  
835 an incontrovertible observation. Instead, Kuhn described scientists as  
836 mostly working within **paradigms**: sets of questions, assumptions, meth-  
837 ods, phenomena, and explanatory hypotheses.

838 Paradigms allow for activities Kuhn described as **normal science**—that is,  
839 testing questions within the paradigm, explaining new observations, or  
840 modifying theory to fit these paradigms. But normal science is punctu-  
841 ated by periods of **crisis** when scientists begin to question their theory  
842 and their methods. Crises don't happen just because a single observa-  
843 tion is inconsistent with the current theory. Rather, there will often be  
844 a holistic transition to a new paradigm, typically because of a striking

845 explanatory or predictive success—often one that's outside the scope of  
 846 the current working theory entirely.

847 In sum, the lesson of holism is that we can't just put our theories in  
 848 direct contact with evidence and think that they will be supported or  
 849 overturned. Instead, we need to think about the scope of our theory (in  
 850 terms of the phenomena and measures it is meant to explain), as well  
 851 as the auxiliary hypotheses—operationalizations—that link it to specific  
 852 observations.

### 853 2.3 Designing experiments to test theory

854 One way of looking at theories is that they let us make *bets*. If we bet  
 855 on a spin of the roulette wheel in figure 2.3 that it will show us red  
 856 as opposed to black, we have almost a 50% chance of winning the bet.

857 Winning such a bet is not impressive. But if we call a particular number,  
 858 the bet is riskier because we have a much smaller chance of being right.

859 Cases where a theory has many chances to be wrong are called **risky tests**  
 860 (Meehl 1978).<sup>7</sup>

861 Much psychology consists of verbal theories. Verbal theories make  
 862 only qualitative predictions, so it is hard convincingly show them to  
 863 be wrong (Meehl 1990). In our discussion of money and happiness,

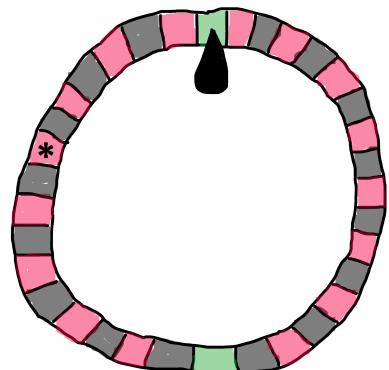


Figure 2.3

A roulette wheel. Betting on red is not that risky, but betting all your chips on a particular value (\*) is much riskier.

<sup>7</sup> Even if you're not a *falsificationist* like Popper, you can still think it's useful to try and falsify theories! Although a single observation is not always enough to overturn a theory, it's still a great research strategy to look for those observations that are most inconsistent with the theory.

864 we just expected happiness to go up as money increased. We would  
865 have accepted *any* increase in happiness (even if very small) as evidence  
866 confirming our hypothesis. Predicting that it does is a bit like betting  
867 on red with the roulette wheel—it's not surprising or impressive when  
868 you win. And in psychology, verbal theories often predict that multiple  
869 factors interact with one another. With these theories, it's easy to say  
870 that one or the other was “dominant” in a particular situation, meaning  
871 you can predict almost any direction of effect.

872 To test theories, we should design experiments to test conditions  
873 where our theories make “risky” predictions. A stronger version of the  
874 money-happiness theory might suggest that happiness increases linearly  
875 in the logarithm of income ([Killingsworth, Kahneman, and Mellers](#)  
876 [2023](#)). This specific mathematical form for the relationship—as well  
877 as the more specific operationalization of money as income—creates  
878 opportunities for making much riskier bets about new experiments.  
879 This kind of case is more akin to betting on a specific number on the  
880 roulette wheel: when you win this bet, it is quite surprising!<sup>18</sup>

881 Testing theoretical predictions also requires precise experimental mea-  
882 surements. As we start to measure the precision of our experimental  
883 estimates in chapter 6, we'll see that the more precise our estimate is,  
884 the more values are inconsistent with it. In this sense, a risky test of a

<sup>18</sup> Theories are often developed iteratively. It's common to start with a theory that is less precise and hence, that has fewer opportunities for risky tests. But by collecting data and testing different alternatives, it's often possible to refine the theory so that it is more specific and allows riskier tests. As we discuss below, formalizing theories using mathematical or computational models is one important route to making more specific predictions and creating riskier tests.

885 theory requires both a very specific prediction and a precise measure-  
886 ment. (Imagine spinning the roulette wheel but seeing such a blurry  
887 image of the result that you can't really tell where the ball is. Not very  
888 useful.)

889 Even when theories make precise predictions, they can still be too flex-  
890 ible to be tested. When a theory has many **free parameters**—numerical  
891 values that can be fit to a particular dataset, changing the theories pre-  
892 dictions on a case-by-case basis—then it can often predict a wide range  
893 of possible results. This kind of flexibility reduces the value of any par-  
894 ticular experimental test, because the theorist can always say after the  
895 fact that the parameters were wrong but not the theory itself (Roberts  
896 and Pashler 2000).

897 One important way to remove this kind of flexibility is to make predic-  
898 tions in advance, holding all parameters constant. A preregistration is a  
899 great way to do this—the experimenter derives predictions and speci-  
900 fies in advance how they will be compared to the results of the exper-  
901 iment. We'll talk much more about the process of preregistration in  
902 chapter 11.

903 We've been focusing mostly on testing a single theory. But the best state  
904 of affairs is if a theory can make a very specific prediction that other  
905 theories don't make. If competing theories both predict that money

906 increases happiness to the same extent, then data consistent with that  
907 predicted relationship don't differentiate between the theories, no mat-  
908 ter how specific the prediction might be. The experiment that teaches  
909 us the most is going to be the one where a very specific pattern of data  
910 is predicted according to one theory and another.<sup>9</sup>

911 Given all of this discussion, as a researcher trying to come up with a  
912 specific research idea, what do you do? Our advice is: *follow the theories*.  
913 That is, for the general topic you're interested in—whether it's money  
914 and happiness, bilingualism, the nature of concepts, or depression—try  
915 to get a good sense of the existing theories. Not all theories will make  
916 specific, testable predictions, but hopefully some will! Then ask, what  
917 are the “risky bets” that these theories make? Do different theories  
918 make different bets about the same effect? If so, that's the effect you  
919 want to measure!

## 920 2.4 Formalizing theories

921 Say we have a set of constructs we want to theorize about. How do we  
922 describe our ideas about the relationships between them so that we can  
923 make precise predictions that can be compared with other theories? As  
924 one writer noted, mathematics is “unreasonably effective” as a vocab-  
925 ular for the sciences (Wigner 1990). Indeed, there have been calls for

<sup>9</sup> We can use this idea, which comes from Bayesian statistics, to try to figure out what the *right* experiment is by considering which specific experimental conditions derive differences between theories. In fact, the idea of choosing experiments based on the predictions that different theories make has a long history in statistics (Lindley 1956); it's now called **optimal experiment design** (Myung, Cavagnaro, and Pitt 2013). The idea is, if you have two or more theories spelled out mathematically or computationally, you can simulate their predictions across a lot of conditions and pick the most informative conditions to run as an actual experiment.

<sup>926</sup> greater formalization of theory in psychology for at least the last 50 years

<sup>927</sup> (Harris 1976).

## DEPTH

### *A universal law of generalization?*

How do you take what you know and apply it to a new situation? One answer is that you use the same answer that has worked in similar situations. To do this kind of extrapolation, however, you need a notion of similarity. Early learning theorists tried to measure similarity by creating an association between a stimulus—say a projected circle of light of a particular size—and a reward by repeatedly presenting them together. After this association was learned, they would test generalization by showing circles of different sizes and measuring the strength of the expectation for a reward. These experiments yielded generalization curves: the more similar the stimulus, the more people and other animals would give the same response, signaling generalization.

Shepard (1987) was interested in unifying the results of these different experiments. The first step in this process was establishing a **stimulus space**. He used a procedure called “multidimensional scaling” to infer how close stimuli were to each other on the basis of how strong the generalization between them was. When he plotted the strength of the generalization by the distance between stimuli within this space (their similarity), he found an incredibly consistent pattern: generalization decreased exponentially as similarity decreased.

He argued that this described a “universal law” that governed the relationship between similarity and generalization for almost any stimulus, whether it was the size of circles, the color of patches of light, or the similarity between speech sounds. Later work has even extended this same framework to highly abstract dimensions such as the relationships between numbers of different types (e.g., being even or being powers of 2; Tenenbaum 1999).

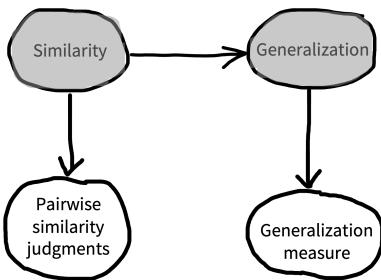


Figure 2.4  
The causal theory of similarity and generalization posited by Shepard (1987).

The pattern shown in Shepard’s work is an example of **inductive theory building**. In the vocabulary we’re developing, Shepard ran (or obtained the data from) randomized experiments in which the manipulation was stimulus dimension (e.g., circle size) and the measure was generalization strength. Then the theory that Shepard proposed was that manipulations of stimulus dimension acted to change the perceived similarity between the stimuli. His theory thus linked two constructs: stimulus similarity and generalization strength (figure 2.4). Critically the causal relationship he described was not just a qualitative relationship but instead a specific mathematical form.

In the conclusion of his paper, Shepard (1987, p. 1323) wrote: “Possibly, behind the diverse behaviors of humans and animals, as behind the various motions of planets and stars, we may discern the operation of universal laws.” While Shepard’s dream is an ambitious one, it defines an ideal for psychological theorizing.

930

931 There is no one approach that will be right for theorizing across all ar-  
932 eas of psychology (Oberauer and Lewandowsky 2019; Smaldino 2020).  
933 Mathematical theories (such as Shepard 1987; see the DEPTH box above)  
934 have long been one tool that allows for precise statements of particular  
935 relationships.

936 Computational or formal artifacts are not themselves psychological the-  
937 ories, but they can be used to create psychological theories via the map-  
938 ping of constructs onto entities in the model and the use of the principles  
939 of the formalism to instantiate psychological hypotheses or assumptions  
940 (Guest and Martin 2021).<sup>10</sup> Yet stating such clear and general laws feels  
941 out of reach in many cases. If we had more Shepard-style theorists or  
942 theories, perhaps we’d be in a better place. Or perhaps such “universal  
943 laws” are simply out of reach for most of human behavior.

944 An alternative approach creates statistical models of data that incorpo-  
945 rate substantive assumptions about the structure of the data. We use  
946 such models all the time for data analysis. The trouble is, we often

947 don't interpret them as having substantive assumptions about the struc-  
948 ture of the data, even when they do (Fried 2020). But if we examine  
949 these assumptions explicitly, even the simplest statistical models can be  
950 productive tools for building theories.

951 For example, if we set up a simple linear regression model to estimate  
952 the relationship between money and happiness, we'd be positing a linear  
953 relationship between the two variables—that an increase in one would  
954 always lead to a proportional increase in the other.<sup>11</sup> If we fit the model  
955 to a particular dataset, we could then look at the weights of the model.  
956 Our theory might then then be something like, “Giving people \$100  
957 causes 0.2 points of increase in happiness on a self-report scale.”

958 Obviously, this regression model is not a very good theory of the broader  
959 relationship between money and happiness, since it posits that every-  
960 one's happiness would be at the maximum on the 10-point scale if you  
961 gave them (at most) \$4,500. It also doesn't tell us how this theory would  
962 generalize to other people, other measures of happiness, or other as-  
963 pects of the psychological representation of money such as income or  
964 wealth.

965 From our viewpoint, these sorts of questions are not distractions—they  
966 are the critical work of moving from experiment to theory (Smaldino  
967 2020)! In chapter 7, we try to draw out this idea further, reconstruing

<sup>11</sup> Linear models are ubiquitous in the social sciences because they are convenient to fit, but as theoretical models they are deeply impoverished. There is a lot you can do with a linear regression, but in the end, most interesting processes are not linear combinations of factors!

968 common statistical tests as models that can be repurposed to express con-  
 969 tentful scientific hypotheses while recognizing the limitations of their  
 970 assumptions.

971 One of the strengths of modern cognitive science is that it provides a  
 972 very rich set of tools for expressing more complex statistical models  
 973 and linking them to data. For example, the modern Bayesian cogni-  
 974 tive modeling tradition grew out of work like Shepard's; in these mod-  
 975 els, a system of equations defines a probability distribution that can be  
 976 used to estimate parameters, predict new data, or make other inferences  
 977 ([Goodman, Tenenbaum, and The ProbMods Contributors 2016](#)). And  
 978 neural network models—which are now fueling innovations in artificial  
 979 intelligence—have a long history of being used as substantive models of  
 980 human psychology ([Elman, Bates, and Johnson 1996](#)). One way to think  
 981 about all these alternatives is as being on a gradient from the general, in-  
 982 spirational frameworks we described above all the way down through  
 983 computational models and then to statistical models that can be fit to  
 984 specific datasets (figure 2.5).

985 In our discussion, we've presented theories as static entities that are  
 986 presented, tested, confirmed, and falsified. That's a simplification that  
 987 doesn't take into account the ways that theories—especially when in-  
 988 stantiated as formal models—can be flexibly adjusted to accommodate

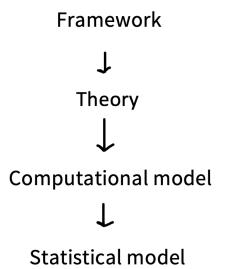


Figure 2.5

A gradient of specificity in theoretical tools. Figure inspired by Guest and Martin (2021).

new data (Navarro 2019). Most modern computational theories are more like a combination of core principles, auxiliary assumptions, and supporting empirical assumptions. The best theories are always being enlarged and refined in response to new data.<sup>12</sup>

### 2.5 Chapter summary: Theories

In this chapter, we characterized psychological theories as a set of causal relationships between latent constructs. The role of experiments is to measure these causal relationships and to adjudicate between theories by identifying cases where different theories make different predictions about particular relationships.

<sup>12</sup> In the thinking of the philosopher Imre Lakatos, a “productive” research program is one where the core principles are gradually supplemented with a limited set of additional assumptions to explain a growing base of observations. In contrast, a “degenerate” research program is one in which you are constantly making ad hoc tweaks to the theory to explain each new datapoint (Lakatos 1976).



#### DISCUSSION QUESTIONS

1. Identify an influential theory in your field or subfield. Can you draw the “nomological network” for it? What are the key constructs and how are they measured? Are the links between constructs just directional links, or is there additional information about what type of relationship exists? Or does our description of a theory in this chapter not fit your example?
2. Can you think of an experiment that falsified a theory in your area of psychology? To what extent is falsification possible for the kinds of theories that you are interested in studying?



## READINGS

- A fabulous introduction to issues in the philosophy of science can be found in: Godfrey-Smith, Peter (2009). *Theory and Reality: An Introduction to the Philosophy of Science*. University of Chicago Press.
- Bayesian modeling has been very influential in cognitive science and neuroscience. A good introduction in cognitive science comes from: Lee, Michael D. and Eric-Jan Wagenmakers (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. Much of the book is available free online at <https://faculty.sites.uci.edu/mdlee/bgm>.
- A recent introduction to Bayesian modeling with a neuroscience focus: Ma, Wei Ji, Konrad Paul Kording, and Daniel Goldreich (2023). *Bayesian Models of Perception and Action: An Introduction*. MIT Press. Free online at <https://www.cns.nyu.edu/malab/bayesianbook.html>.

1000

<sup>1001</sup> *References*

- Block, Ned J., and Jerry A. Fodor. 1972. “What Psychological States Are Not.” *The Philosophical Review* 81 (2): 159–181.
- Bronfenbrenner, Urie. 1992. *Ecological Systems Theory*. Jessica Kingsley Publishers.
- Clarke, Edward H. 1884. *Sex in Education: Or, a Fair Chance for the Girl*. Boston: Houghton Mifflin.
- <sup>1002</sup> Cronbach, Lee J., and Paul E. Meehl. 1955. “Construct Validity in Psycholog-

- ical Tests.” *Psychological Bulletin* 52 (4): 281–302.
- Ellwood-Lowe, Monica E., Ruthe Foushee, and Mahesh Srinivasan. 2022. “What Causes the Word Gap? Financial Concerns May Systematically Suppress Child-Directed Speech.” *Developmental Science* 25 (1): e13151.
- Elman, Jeffrey L., Elizabeth A. Bates, and Mark H. Johnson. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Vol. 10. MIT Press.
- Fried, Eiko I. 2020. “Lack of Theory Building and Testing Impedes Progress in the Factor and Network Literature.” *Psychological Inquiry* 31 (4): 271–288.
- Godfrey-Smith, Peter. 2009. *Theory and Reality: An Introduction to the Philosophy of Science*. University of Chicago Press.
- Goodman, Noah D., Joshua B. Tenenbaum, and The ProbMods Contributors. 2016. “Probabilistic Models of Cognition.” <http://probmods.org/>.
- Guest, Olivia, and Andrea E. Martin. 2021. “How Computational Modeling Can Force Theory Building in Psychological Science.” *Perspectives on Psychological Science* 16 (4): 789–802.
- Harris, Richard J. 1976. “The Uncertain Connection between Verbal Theories and Research Hypotheses in Social Psychology.” *Journal of Experimental Social Psychology* 12 (2): 210–219.
- Hempel, Carl G. 1952. *Fundamentals of Concept Formation in Empirical Science*. University of Chicago Press.
- Killingsworth, Matthew A., Daniel Kahneman, and Barbara Mellers. 2023. “Income and Emotional Well-Being: A Conflict Resolved.” *Proceedings of the National Academy of Sciences* 120 (10): e2208661120.
- Kraus, Michael W., Ivuoma N. Onyeador, Natalie M. Daumeyer, Julian M. Rucker, and Jennifer A. Richeson. 2019. “The Misperception of Racial

- Economic Inequality.” *Perspectives on Psychological Science* 14 (6): 899–921.
- Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Princeton University Press.
- Lakatos, Imre. 1976. “Falsification and the Methodology of Scientific Research Programmes.” In *Can Theories Be Refuted?*, 205–259. Springer.
- Lee, Michael D., and Eric-Jan Wagenmakers. 2013. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.
- Lindley, Dennis V. 1956. “On a Measure of the Information Provided by an Experiment.” *The Annals of Mathematical Statistics*, 986–1005.
- Ma, Wei Ji, Konrad Paul Kording, and Daniel Goldreich. 2023. *Bayesian Models of Perception and Action: An Introduction*. MIT Press.
- Meehl, Paul E. 1978. “Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology.” *Journal of Consulting and Clinical Psychology* 46 (4): 806–834.
- Meehl, Paul E. 1990. “Why Summaries of Research on Psychological Theories Are Often Uninterpretable.” *Psychological Reports* 66 (1): 195–244.
- Myung, Jay I., Daniel R. Cavagnaro, and Mark A. Pitt. 2013. “A Tutorial on Adaptive Design Optimization.” *Journal of Mathematical Psychology* 57 (3–4): 53–67.
- Navarro, Danielle J. 2019. “Between the Devil and the Deep Blue Sea: Tensions between Scientific Judgement and Statistical Model Selection.” *Computational Brain & Behavior* 2 (1): 28–34.
- Oberauer, Klaus, and Stephan Lewandowsky. 2019. “Addressing the Theory Crisis in Psychology.” *Psychonomic Bulletin & Review* 26 (5): 1596–1618.
- Roberts, Seth, and Harold Pashler. 2000. “How Persuasive Is a Good Fit? A

- Comment on Theory Testing.” *Psychological Review* 107 (2): 358.
- Rohrer, Julia M. 2018. “Thinking Clearly about Correlations and Causation: Graphical Causal Models for Observational Data.” *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42.
- Sandvik, Ed, Ed Diener, and Larry Seidlitz. 1993. “Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures.” *Journal of Personality* 61 (3): 317–342.
- Shepard, Roger N. 1987. “Toward a Universal Law of Generalization for Psychological Science.” *Science* 237 (4820): 1317–1323.
- Smaldino, Paul E. 2020. “How to Translate a Verbal Theory into a Formal Model.” *Social Psychology* 51 (4): 207–218.
- Sprenger, Jan, and Stephan Hartmann. 2019. *Bayesian Philosophy of Science*. Oxford University Press.
- Strevens, Michael. 2006. “The Bayesian Approach to the Philosophy of Science.” In *Encyclopedia of Philosophy, Second Edition*, edited by D. M. Borchert, 495–502. Macmillan Reference.
- Tenenbaum, Joshua B. 1999. “Rules and Similarity in Concept Learning.” In *Advances in Neural Information Processing Systems*, 12:59–65. MIT Press.
- Ungar, Michael. 2002. “A Deeper, More Social Ecological Social Work Practice.” *Social Service Review* 76 (3): 480–497.
- Wigner, Eugene P. 1990. “The Unreasonable Effectiveness of Mathematics in the Natural Sciences.” In *Mathematics and Science*, 291–306. World Scientific.

## <sup>1006</sup> 3 REPLICATION

### LEARNING GOALS

- Define and distinguish reproducibility and replicability
- Synthesize the metascientific literature on replication and the causes of replication failures
- Reason about the relation of replication to theory building

<sup>1007</sup>

<sup>1008</sup> In the previous chapters, we introduced experiments, their connection  
<sup>1009</sup> with causal inference, and their role in building psychological theory. In  
<sup>1010</sup> principle, repeated experimental work combined with theory building  
<sup>1011</sup> should yield strong research programs that explain and predict phenom-  
<sup>1012</sup> ena with increasing scope.

<sup>1013</sup> Yet, in recent years there has been an increasing recognition that this ide-  
<sup>1014</sup> alized view of science might not be a good description of what we actu-  
<sup>1015</sup> ally see when we look at the psychology literature. Many classic findings  
<sup>1016</sup> may be wrong, or at least overstated. Their statistical tests might not be

1017 trustworthy. The actual numbers are even wrong in many papers! And  
 1018 even when experimental findings are “real,” they may not generalize  
 1019 broadly to different people and different situations.

1020 How do we know about these problems? A burgeoning field called  
 1021 **metascience** is providing the evidence. Metascience is research *about*  
 1022 *research*—for example, investigating how often findings in a literature  
 1023 can be successfully built on, or trying to figure out how widespread  
 1024 some negative practice is. Metascience allows us to go beyond one-off  
 1025 anecdotes about a particular set of flawed results or rumors about bad  
 1026 practices. Perhaps the most obvious sign that something is wrong is that  
 1027 when independent scientists team up in metascience projects and try to  
 1028 repeat previous studies, they often do not get the same results.

1029 Before we begin reviewing this evidence, let’s discuss the different ways  
 1030 in which a scientific finding can be repeated. Figure 3.1 gives us a basic  
 1031 starting point for our definitions. For a particular finding in a paper, if  
 1032 we take the same data, do the same analysis, and get the same result, we  
 1033 call that finding **reproducible** (sometimes, **analytically** or **computation-**  
 1034 **ally reproducible**). If we collect *new* data using the same methods, do  
 1035 the same analysis, and get the same result, we call that a **replication** and  
 1036 say that the finding is **replicable**. If we do a different analysis with the  
 1037 same data, we call this a **robustness check** and if we get a similar result,

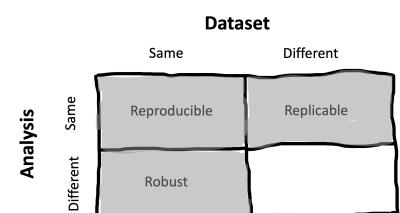


Figure 3.1  
 A framework for understanding different terms related to the repeatability of scientific findings (based on Whitaker 2017).

1038 we say that the finding is **robust**.<sup>1</sup> We leave the last quadrant empty be-  
1039 cause there's no specific term for it in the literature—the eventual goal  
1040 is to draw **generalizable** conclusions, but this term means more than just  
1041 having a finding that is reproducible and replicable.

1042 In this chapter, we'll primarily discuss reproducibility and replicability  
1043 (we'll talk about robustness a bit in chapter 11). We'll start out by re-  
1044 viewing key concepts around reproducibility and replicability as well  
1045 as some important metascience findings. This literature suggests that  
1046 when you read an average psychology paper, your default expectation  
1047 should be that it might not replicate!

1048 We'll then discuss some of the main reasons *why* findings might not  
1049 replicate—especially **analytic flexibility** and **publication bias**. We end  
1050 by taking up the issue of how reproducibility and replicability relate to  
1051 theory building in psychology, and the role of **open science** in this dis-  
1052 cussion. This discussion focuses on the key role of **TRANSPARENCY** (one  
1053 of our major book themes) in supporting theory building.

### 1054 3.1 Reproducibility

1055 Scientific papers are full of numbers: sample sizes, measurements, statis-  
1056 tical results, and visualizations. For those numbers to have meaning, and

<sup>1</sup> You might have observed that a lot of work is being done here by the word “same.” How do we operationalize same-ness for experimental procedures, statistical analyses, samples, or results? These are difficult questions that we'll touch on below. Keep in mind that there's no single answer and so these terms are always going to helpful guides rather than exact labels.

1057 for other scientists to be able to verify them, we need to know where  
1058 they came from (their *provenance*). The chain of actions that scien-  
1059 tists perform on the raw data, all the way through to reporting numbers  
1060 in their papers, is sometimes called the *analysis pipeline*. For much of  
1061 history, scientific papers have only provided a verbal description of the  
1062 analysis pipeline, usually with little detail.<sup>2</sup>

1063 Moreover, researchers typically do not share key research objects from  
1064 this pipeline, such as the analysis scripts or the raw data (Hardwicke,  
1065 Thibault, et al. 2021).<sup>3</sup> Without code and data, the numbers reported  
1066 in scientific papers are often not reproducible—an independent scientist  
1067 cannot repeat all of the steps in the analysis pipeline and get the same  
1068 results as the original scientists.

1069 Reproducibility is desirable for a number of reasons. Without it:

- 1070 – Errors in calculation or reporting could lead to disparities  
1071 between the reported result and the actual result
- 1072 – Vague verbal descriptions of analytic computations could keep  
1073 readers from understanding the computations that were actually  
1074 performed
- 1075 – The robustness of data analyses to alternative model specifications  
1076 cannot be checked

<sup>2</sup> The situation is nicely summed up by a prescient quote from Buckheit and Donoho (1995, p. 5): “A scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”.

<sup>3</sup> For many years, professional societies, like the American Psychological Association, have mandated data sharing (<https://www.apa.org/ethics/code>), but only for purposes of verification, and only “on request”—in other words, scientists could keep data hidden by default and it was their responsibility to share if another scientist requested access. In practice, this kind of policy does not work; data are rarely made available on request (Wicherts et al. 2006). We believe this situation is untenable. We provide a longer argument justifying data sharing in chapter 4 and discuss some of the practicalities of sharing in chapter 13.

1077 – Synthesizing evidence across studies, a key part of building a cu-  
1078 mulative body of scientific knowledge, is much more difficult

1079 From this list, error detection and correction is probably the most press-  
1080 ing issue. But are errors common? There are plenty of individual in-  
1081 stances of errors that are corrected in the published literature (e.g., some  
1082 of us found an error in [Cesana-Arlotti et al. 2018](#)), and we ourselves  
1083 have made significant analytic errors (e.g., [Frank et al. 2013](#)). But  
1084 these kinds of experiences don't tell us about the frequency of errors  
1085 more generally (or the consequences of error for the conclusions that  
1086 researchers draw).<sup>4</sup>

1087 Estimating the frequency of errors is a metascientific issue that re-  
1088 searchers have attempted to answer over the years. If errors are  
1089 frequent, that would suggest a need for changes in our policies and  
1090 practices to reduce their frequency! Unfortunately, the lack of data  
1091 availability creates a problem: it's hard to figure out if calculations are  
1092 wrong if you can't check them in the first place. Here's one clever  
1093 approach to this issue. In standard American Psychological Association  
1094 (APA) reporting format, inferential statistics must be reported with  
1095 three pieces of information: the test statistic, the degrees of freedom  
1096 for the test, and the  $p$ -value (e.g.,  $t(18) = -0.74, p = 0.47$ ). Yet,  
1097 these pieces of information are redundant with one another. Thus,

<sup>4</sup> There is a very interesting discussion of the pernicious role of scientific error on theory building in Gould's ([1996](#)) "The Mismeasure of Man." Gould examines research on racial differences in intelligence and documents how scientific errors that supported racial differences were often overlooked. Errors are often caught asymmetrically; we are more motivated to double-check a result that contradicts our biases.

1098 reported statistics can be checked for consistency simply by evaluating  
1099 whether they line up with one another—that is, whether the  $p$ -value  
1100 recomputed from the  $t$  and degrees of freedom matches the reported  
1101 value.

1102 Bakker and Wicherts (2011) performed this kind of statistical consis-  
1103 tency analysis on a sample of 281 papers, and found that around 18%  
1104 of statistical results were incorrectly reported. Even more worrisome,  
1105 around 15% of articles contained at least one decision error—that is, a  
1106 case where the error changed the direction of the inference that was  
1107 made (e.g., from significant to insignificant).<sup>5</sup> Nuijten et al. (2016) used  
1108 an automated method called “statcheck”<sup>6</sup> to confirm and extend this  
1109 analysis. They checked  $p$ -values for more than 250,000 psychology pa-  
1110 pers in the period 1985–2013 and found that around half of all papers  
1111 contained at least one incorrect  $p$ -value!

1112 These findings provide a lower bound on the number of errors in the  
1113 literature and suggest that reproducibility of analyses is likely very im-  
1114 portant. However, they only address the consistency of statistical re-  
1115 porting. What would happen if we tried to repeat the entire analysis  
1116 pipeline from start to finish? It’s rather difficult to answer this question  
1117 at a large scale: firstly, it takes a long time to run a reproducibility check;  
1118 and secondly, the lack of access to raw data means that for most scientific

<sup>5</sup> Confirming Gould’s speculation (see note above), most of the reporting errors that led to decision errors were in line with the researchers’ own hypotheses.

<sup>6</sup> Statcheck is now available as a web app (<http://statcheck.io>) and an R package (Nuijten and Epskamp 2024) so that you can check your own manuscripts!

<sup>1119</sup> papers, checking reproducibility is impossible.

<sup>1120</sup> Nevertheless, a few years ago a group of us spotted an opportunity  
<sup>1121</sup> to check reproducibility by examining studies published in two jour-  
<sup>1122</sup> nals that either required or encouraged data sharing. Hardwicke et al.  
<sup>1123</sup> (2018) and Hardwicke, Bohn, et al. (2021) first identified studies that  
<sup>1124</sup> shared data, then narrowed those down to studies that shared *reusable*  
<sup>1125</sup> data (the data were accessible, complete, and comprehensible). For 60  
<sup>1126</sup> of these articles, we then attempted to reproduce numerical values re-  
<sup>1127</sup> lated to a particular statistical result in the paper. The process was in-  
<sup>1128</sup> credibly labor-intensive, with articles typically requiring 5–10 hours  
<sup>1129</sup> of work each. And the results were concerning: the targeted values in  
<sup>1130</sup> only about a third of articles were completely reproducible without help  
<sup>1131</sup> from the original authors! In many cases, after—sometimes extensive—  
<sup>1132</sup> correspondence with the original authors, they provided additional in-  
<sup>1133</sup> formation that was not reported in the original paper. After author con-  
<sup>1134</sup> tact, the reproducibility success rate improved to 62% (figure 3.2). The  
<sup>1135</sup> remaining papers appeared to have some values that neither we, nor  
<sup>1136</sup> the original authors, could reproduce. Importantly, we didn’t identify  
<sup>1137</sup> any patterns of nonreproducibility that seriously undermined the con-  
<sup>1138</sup> clusions drawn in the original articles; however, other reproducibility  
<sup>1139</sup> studies have found a distressingly high number of decision errors (Artner  
<sup>1140</sup> et al. 2020), albeit with a slightly higher success rate overall.

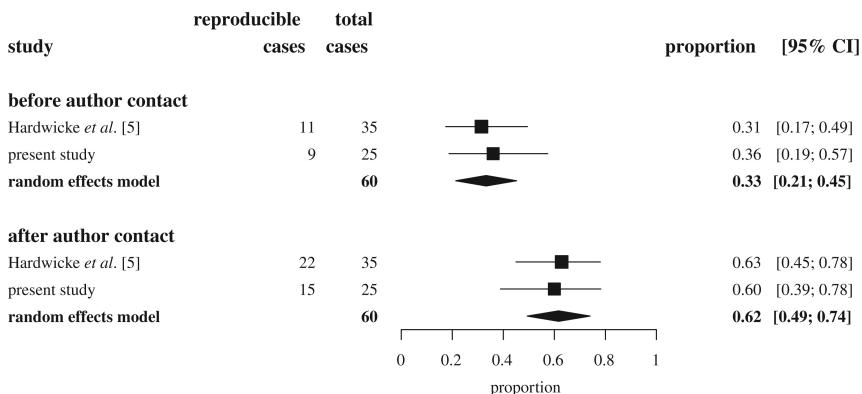


Figure 3.2

Analytic reproducibility of results from open-data articles in *Cognition* and *Psychological Science*. From Hardwicke, Bohn, et al. (2021), figure 1 (licensed under CC BY 4.0).

- 1141 In sum: transparency is a critical imperative for decreasing the fre-  
 1142 quency of errors in the published literature. Reporting and computation  
 1143 errors are frequent in the published literature, and the identification of  
 1144 these errors depends on the findings being reproducible. If data are not  
 1145 available, then errors usually cannot be found.

### CASE STUDY

#### *The Open Science Collaboration*

Around 2011, we were teaching our Experimental Methods course for the first time, based on a course model that we had worked on with Rebecca Saxe (Frank and Saxe 2012). The idea was to introduce students to the nuts and bolts of research by having them run replications. A guy named Brian Nosek was on sabbatical nearby, and over coffee we learned that he was starting up an ambitious project to replicate a large sample of studies published in top psychology journals in 2008.

In the course that year we chose replication projects from the sample that Nosek had told us about. Four of these projects were executed very

well and were nominated by the course TAs for inclusion in the broader project. A few years later, when the final group of 100 replication studies was completed, we got a look at the results, shown in figure 3.3.

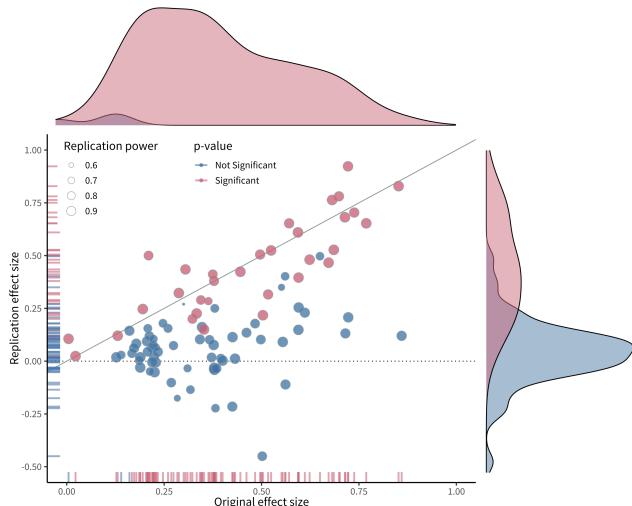


Figure 3.3  
Results from the Open Science Collaboration (2015). Each point represents one of the studies in the sample, with the horizontal position giving the original effect size and the vertical position giving the replication effect size. Dot size represents estimated statistical power. The grey line represents a perfect replication.

The resulting metascience paper, which we and others refer to as the “replication project in psychology” (RPP), made a substantial impression on both psychologists and the broader research community, defining both a field of psychology metascience studies and providing a template for many-author collaborative projects (Open Science Collaboration 2015). But the most striking thing was the result: disappointingly, only around a third of the replications had similar findings to the original studies. The others yielded smaller effects that were not statistically significant in the replication sample (almost all of the original studies were significant). The RPP provided the first large-scale evidence that there were systematic is-

sues with replicability in the psychology literature.

The RPP's results—and their interpretation—were controversial, however, and much ink was spilled on what these data showed. In particular, critics pointed to different degrees of fidelity between the original studies and the replications; insufficient levels of statistical power and low evidential value in the replications; nonrepresentative sampling of the literature; and difficulties identifying specific statistical outcomes for replication success (Gilbert et al. 2016; Anderson et al. 2016; Etz and Vandekerckhove 2016). In our view, many of these critiques have merit, and you can't simply interpret the results of RPP as an unbiased estimate of the replicability of results in the literature, contra the title.

And yet, RPP's results are still important and compelling, and they undeniably changed the direction of the field of psychology. Many good studies are like this—they have flaws but they inspire follow-up studies that can address those problems. For several of us personally, working on this project was also transformative in that it showed us the power of collaborative work. Together we could do a study that no one of us had any hope of completing on our own, and potentially make a difference in our field.

1148

### *3.2 Replication*

1149 Beyond verifying a paper's original analysis pipeline, we are often interested in understanding whether the study can be replicated—if we

1152 repeat the study methods and obtain new data, do we get similar re-  
1153 sults? To quote from Popper (1959, p. 86), “The scientifically signifi-  
1154 cant … effect may be defined as that which can be regularly [replicated]  
1155 by anyone who carries out the appropriate experiment in the way pre-  
1156 scribed.”

1157 Replications can be conducted for many reasons (Schmidt 2009). One  
1158 goal can be to verify that the results of an existing study can be obtained  
1159 again if the study is conducted again in exactly the same way, to the  
1160 best of our abilities. A second goal can be to gain a more precise esti-  
1161 mate of the effect of interest by conducting a larger replication study, or  
1162 combining the results of a replication study with the existing study. A  
1163 third goal can be to investigate whether an effect will persist when, for  
1164 example, the experimental manipulation is done in a different, but still  
1165 theory-consistent, manner. Alternatively, we might want to investigate  
1166 whether the effect persists in a different population. Such replications  
1167 are often efforts to “replicate and extend,” and are common both when  
1168 the same research team wants to conduct a sequence of experiments  
1169 that each build on one another or when a new team wants to build on  
1170 a result from a paper they have read (Rosenthal 1990).

1171 Much of the metascience literature (and attendant debate and discus-  
1172 sion) has focused on the first goal—simple verification. This focus has

1173 been so intense that the term “replication” has become associated with  
1174 skepticism or even attacks on the foundations of the field. This dynamic  
1175 is at odds with the role that replication is given in a lot of philosophy of  
1176 science, where it is assumed to be a typical part of “normal science.”

1177 *3.2.1 Conceptual frameworks for replication*

1178 The key challenge of replication is **invariance**—Popper’s stipulation that  
1179 a replication be conducted “in the way prescribed” in the quote above.  
1180 That is, what are the features of the world over which a particular obser-  
1181 vation should be relatively constant, and what are those that are specified  
1182 as the key ingredients for the effect? Replication is relatively straightfor-  
1183 ward in the physical and biological sciences, in part because of presup-  
1184 posed theoretical background that allows us to make strong inferences  
1185 about invariance. If a biologist reports an observation about a particular  
1186 cell type from an organism, the color of the microscope is presumed not  
1187 to matter to the observation.

1188 These invariances are far harder to state in psychology, for both the pro-  
1189 cedure of an experiment and its sample. Procedurally, should the color  
1190 of the experimental stimulus matter to the measured effect? In some  
1191 cases yes, in some cases no.<sup>7</sup> Yet, the task of postulating how a scientific  
1192 effect should be invariant to lab procedures pales in comparison to the

<sup>7</sup> A fascinating study by Baribault et al. (2018) proposes a method for em-  
pirically understanding psychological in-  
variances. Treating a subliminal priming  
effect as their model system, they sam-  
pled thousands of “micro-experiments”  
in which small parameters of their exper-  
imental procedure were randomly sam-  
pled. These parameters allowed for mea-  
surement of their effect of interest, aver-  
aging across this irrelevant variation. In  
their case, it turned out that color did not  
matter.

1193 task of postulating how the effect should be invariant across different  
1194 human populations!<sup>8</sup>

1195 A lot is at stake in this discussion. If Dr. Frog publishes a finding with  
1196 US undergraduates and Dr. Toad then “replicates” the procedure in  
1197 Germany, to what extent should we be perturbed if the effect is differ-  
1198 ent in magnitude or absent?<sup>9</sup> Meta-researchers have made a number of  
1199 replication taxonomies to try and quantify the degree of methodologi-  
1200 cal consistency between two experiments.

1201 Some researchers have tried to distinguish “direct replications”<sup>10</sup> and  
1202 “conceptual replications.” Direct replications are those that attempt to  
1203 reproduce all of the salient features of the prior study, up to whatever in-  
1204 variances the experimenters believe are present (e.g., color of the paint  
1205 or gender of the experimenter). In contrast, conceptual replications are  
1206 typically paradigms that attempt to test the same hypothesis via differ-  
1207 ent operationalizations of the manipulation and/or the measure. We  
1208 agree with Zwaan et al. (2018): labeling this second type of experiment  
1209 as a “replication” is a little misleading. Rather, so-called conceptual  
1210 replications are actually different tests of the same part of your theory.  
1211 Such tests can be extremely valuable, but they serve a different goal than  
1212 replication.

<sup>8</sup> In some sense, the research program of some branches of the social sciences amounts to an understanding of invariances across human cognition.

<sup>9</sup> Presumably not very much if Dr. Toad gave the original instructions in English instead of in German—that’s another one of these pesky invariances that we are always worrying about!

<sup>10</sup> These also get called **exact** replica-  
tions sometimes. We think this term is misleading because similarity between two different experiments is always going to be on a gradient, and where you cut this continuum is always going to be a theory-laden decision. One person’s “exact” is another’s “inexact.”

## 💥 ACCIDENT REPORT

### *“Small Telescopes”*

We've been discussing the question of invariance with respect to procedure and sample, but we haven't really discussed invariance with respect to the studies' statistical results. To what extent can we consider two statistical results to be “the same”? Several obvious metrics, including those used by RPP, have important limitations (Simonsohn 2015). For example, if one finding is statistically significant and the other isn't, they still could have effect sizes that are actually quite close to one another, in part because one might have a larger sample size than the other. Or you could have two significant findings that nevertheless have very different effect sizes.

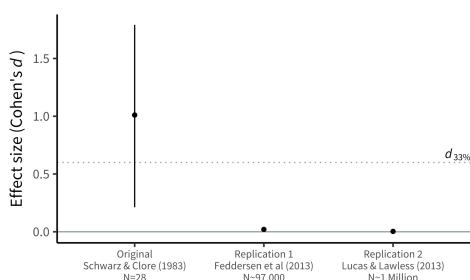


Figure 3.4

The original finding by Schwarz and Clore (1983) and two replications with much larger samples. All three estimates include a 95% confidence interval, but the confidence intervals are very small for the two replication studies. The blue dotted line shows the smallest effect that the original study could reasonably have detected. Based on Simonsohn (2015).

In a classic study, Schwarz and Clore (1983) reported that participants ( $N = 28$ ) rated their life satisfaction as higher on sunny days than rainy days, suggesting that they misattributed temporary happiness about the

weather to longer-term life satisfaction. However, when two more recent studies examined very large samples of survey responses, they yielded estimates of the effect that were much smaller. (All of these effects have been standardized so that they are on the same scale using a metric called Cohen's  $d$  that we will introduce more formally in chapter 5.) In one survey, the effect was statistically significant but extremely small; in the other it was essentially zero (figure 3.4). Using statistical significance as the metric of replication success, you might be tempted to say that the first of these studies was a successful replication and the second was a failed replication.

Simonsohn points out that this interpretation doesn't make sense, using the analogy of a study's sample size as a telescope. Following this analogy, Schwarz and Clore had a very small telescope (i.e., a small sample size), and they pointed it in a particular direction and claimed to have observed a planet (i.e., a nonzero effect). Now it might turn out that there *was* a planet at that location when you look with a much larger telescope (first replication), and it might turn out that there *wasn't* (second replication). Regardless, however, the original small telescope was simply not powerful enough to have seen whatever was there. Both studies fail to replicate the original observation, regardless of whether their observed effect was in the same direction.

Following Simonsohn's example, numerous metrics for replication success have been proposed (Mathur and VanderWeele 2020). The best of these move away from the idea that there is a binary test of whether an individual replication was successful and toward a comparison of the two

effects and whether they appear consistent with the same theory. Gelman (2018) suggests the “time reversal” heuristic—rather than thinking of a replication as a success or a failure, consider the alternative world in which the replication study had been performed first and the original study followed it.

If we leave behind the idea that the original study has precedence, it makes much more sense to consider the sum total of the evidence across multiple experiments. Using this approach, it seems pretty clear that the weather misattribution effect is, at best, a tiny factor in people’s overall judgments of their life satisfaction, even if a small study once found a larger effect.

1215

### 1216 3.2.1 *The metascience of replication*

1217 In RPP, replication teams reported subjectively that 39% of replications  
1218 were successful, with 36% reporting a significant effect in the same di-  
1219 rection as the original. How generalizable is this estimate—and how  
1220 replicable *is* psychological research more broadly? Based on the discus-  
1221 sion above, we hope we’ve made you skeptical that this is a well-posed  
1222 question, at least without additional qualifiers. Any answer is going to  
1223 have to provide details about the scope of this claim, the definition of  
1224 replication being used, and the metric for replication success. On the  
1225 other hand, *versions* of this question have led to a number of empirical

1226 studies that help us better understand the scope of replication issues.

1227 Many subsequent empirical studies of replication have focused on par-  
1228 ticular subfields or journals, with the goal of informing particular field-  
1229 specific practices or questions. For example, Camerer et al. (2016) repli-  
1230 cated all of the between-subject laboratory articles published in two top  
1231 economics journals in the period 2011–2014. They found a replication  
1232 rate of 61% of significant effects in the same direction of the original,  
1233 higher than the rate in RPP but lower than the naive expectation based  
1234 on their level of statistical power. Another study attempted to replicate  
1235 all 21 behavioral experiments published in the journals *Science* and *Na-*  
1236 *ture* in 2010–2015, finding a replication rate of 62% significant effects  
1237 (Camerer et al. 2018). This study was notable because they followed  
1238 a two-step procedure—after an initial round of replications, they fol-  
1239 lowed up on the failures by consulting with the original authors and  
1240 pursuing extremely large sample sizes. The resulting estimate thus is  
1241 less subject to many of the critiques of the original RPP paper. While  
1242 these types of studies do not answer all the questions that were raised  
1243 about RPP, they suggest that replication rates for top experiments are  
1244 not as high as we’d like them to be, even when care is taken with the  
1245 sampling and individual study protocols.

1246 Other scientists working in the same field can often predict when an ex-

periment will fail to replicate. Dreber et al. (2015) showed that prediction markets (where participants bet small sums of real money on replication outcomes) made fairly accurate estimates of replication success in the aggregate. This result has itself now been replicated several times (e.g., in the Camerer et al. 2018 study described earlier). Maybe even more surprisingly, there's some evidence that machine learning models trained on the text of papers can predict replication success (Yang, Youyou, and Uzzi 2020; Youyou, Yang, and Uzzi 2023), though more work still needs to be done to validate these models and understand the features they use. More generally, these two lines of research suggest the possibility of isolating consistent factors that lead to replication success or failure. (In the next section we consider what these factors are in more depth.)

Although more work still needs to be done to get generalizable estimates of replicability, taken together, the metascience literature does provide some clarity on what we should expect. Altogether, the chance of a significant finding in a (well-powered) replication study of a generic experiment in social and cognitive psychology is likely somewhere around 56%. Furthermore, the replication effect will likely be on average 53% as large (Nosek et al. 2022).

On the other hand, these large-scale replication studies have substantial

1268 limitations as well. With relatively few exceptions, the studies chosen  
1269 for replication used short, computerized tasks that mostly would fall  
1270 into the categories of social and cognitive psychology. Further, and per-  
1271 haps most troubling from the perspective of theory development, they  
1272 tell us only whether a particular experimental effect can be replicated.  
1273 They tell us much less about whether the construct that the effect was  
1274 meant to operationalize is in fact real! We'll return to the difficult issue  
1275 of how replication and theory construction relate to one another in the  
1276 final section of this chapter.

1277 Some have called the narrative that emerges from the sum of these meta-  
1278 science studies the “replication crisis.” We think of it as a major temper-  
1279 ing of expectations with respect to the published literature. Your naive  
1280 expectation might reasonably be that you could read a typical journal ar-  
1281 ticle, select an experiment from it, and replicate that experiment in your  
1282 own research. The upshot of this literature is, unfortunately, if you try  
1283 selecting and replicating an exeriment, you might well be disappointed  
1284 by the result.

#### INCIDENT REPORT

##### *Consequences for the study, consequences for the person*

“Power posing” is the idea that adopting a more open and expansive phys-  
ical posture might also change your confidence. Carney, Cuddy, and Yap

(2010) told 42 participants that they were taking part in a study of physiological recording. They then held two poses, each for a minute. In one condition, the poses were expansive (e.g., legs out, hands on head); in another condition, the poses were contractive (e.g., arms and legs crossed). Participants in the expansive pose condition showed increases in testosterone and decreases in salivary cortisol (a stress marker), they took a greater number of risk in a gambling task, and they reported that they were more “in charge” in a survey. This result suggested that a two-minute manipulation could lead to striking physiological and psychological changes—in turn leading to power posing becoming firmly enshrined as part of the set of recommended strategies in business and elsewhere. The original publication contributed to the rise of the researchers’ careers, including becoming a principal piece of evidence in a hugely popular TED talk by Amy Cuddy, one of the authors.

Follow-up work has questioned these findings, however. A replication study with a larger number of participants ( $N = 200$ ) failed to find evidence for physiological effects of power posing, even as it did find some effects on participants’ own beliefs (Ranehill et al. 2015). And a review of the published literature suggested that many findings appeared to be the result of some sort of publication bias, as far too many of them had  $p$ -values very close to the 0.05 threshold (Simmons and Simonsohn 2017). In light of this evidence, the first author of the replication study bravely made a public statement that she does not believe that “power pose” effects are real (Carney 2016).

From the scientific perspective, it's very tempting to take this example as a case in which the scientific ecosystem corrects itself. Although many people continue to cite the original power posing work, we suspect the issues are well known throughout the social psychology community, and overall interest from the lay public has gone down. But this narrative masks the very real human impacts of the self-correction process, which can raise ethical questions about the best way to address issues in the scientific record.

The process of debate and discussion around individual findings can be bruising and complicated. In the case of power posing, Cuddy herself was tightly associated with the findings, and many critiques of the findings became critiques of the individual. Several commentators used Cuddy's name as a stand-in for low-quality psychological results, likely because of her prominence and perhaps because of her gender and age as well. These comments were harmful to Cuddy personally and her career more generally.

Scientists should critique, reproduce, and replicate results—these are all parts of the progress of normal science. But it's important to do this in a way that's sensitive to the people involved. Here are a few guidelines for courteous and ethical conduct:

- Always communicate about the work, never the person. Try to use language that is specific to the analysis or design being critiqued, rather than the person who did the analysis or thought up the design.

- Avoid using language that assumes negative intentions, for example, “the authors misleadingly state that …”
- Ask someone to read your paper, email, blogpost, or tweet before you hit send. It can be very difficult to predict how someone else will experience the tone of your writing; a reader can help you make this judgement.
- Consider communicating personally before communicating publicly. As Joe Simmons, one critic in the power posing debate, said, “I wish I’d had the presence of mind to pick up the phone and call [before publishing my critique]” (Dominus 2017). Personal communication isn’t always necessary (and can be difficult due to asymmetries of power or status), but it can be helpful.

As we will argue in the next chapter, we have an ethical duty as scientists to promote good science and critique low-quality science. But we also have a duty to our colleagues and communities to be good to one another.

1288

### *3.3 Causes of replication failure*

#### ⊕ DEPTH

##### *Context, moderators, and expertise*

There are many explanations for failed replications. The wonderful thing about metascience is that these explanations can be tested empirically!

Let’s start with the idea that specific experimental operationalizations of a

1290

theory might be “context sensitive,” especially in subfields, like social psychology, whose theories inherently refer to environmental context (Van Bavel et al. 2016). Critics brought this issue up for RPP, where there were several studies in which the original experimental materials were tailored to one cultural context but then were deployed in another context, potentially leading to failure due to mismatch (Gilbert et al. 2016).

Context sensitivity seems like a great explanation because in some sense, it *must* be right. If the context of an experiment includes the vast network of learned associations, practices, and beliefs that we all hold, then there’s no question that an experiment’s materials tap into this context to one degree or another. For example, if your experiment relies on the association between *doctor* and *nurse* concepts, you would expect this experiment to work differently in the past when *nurse* meant something more like *nanny* (Ramscar 2016).

On the other hand, as an explanation of specific replication failures, context sensitivity has not fared very well. The “Many Labs” projects were a series of replication projects in which *multiple* labs independently attempted to replicate several original studies. (In contrast, in RPP and similar studies, a single replication was conducted for each original study.) Some of the Many Labs projects assessed variation in replication success across different labs. In ManyLabs 2, Klein et al. (2018) replicated 28 findings, distributed across 125 different samples and more than 15,000 participants. ManyLabs 2 found almost no support for the context sensitivity hypothesis as an explanation of replication failure. In general, when

effects failed to replicate, they did so when conducted in person as well as when conducted online, and these failures were consistent across many cultures and labs.

On the other hand, a review of several Many Labs–style replication projects indicated, on reanalysis, that population effects differed across replication labs even when the replication protocols were very similar to one another (Olsson-Collentine, Wicherts, and Assen 2020; Errington et al. 2021). So, context sensitivity is almost certainly present—and we’ll return to the broader issues of generalizability, context, and invariance in the next section—but so far we have not identified specific forms of context sensitivity that reliably affect replication success.

These observations—that (1) direct replications vary in how successful they are, but (2) we cannot identify specific contextual moderators—together suggest the possible presence of “hidden moderators.” That is, when faced with a successful original study and a failed replication, there may be some unknown factor(s) that moderates the effect.

We’ve personally had several experiences that corroborate the idea that there are hidden moderators. For example, in Lewis and Frank (2016), we were unsuccessful in replicating a simple categorization experiment. We then made a series of iterative changes to the stimuli and instructions, for example, changing the color and pattern of the stimuli (figure 3.5), eventually resulting in a larger (and statistically significant) effect—though still much smaller than the original. Critically, however, each alteration that

we made to the procedure yielded a very small change in the effect, and it would have taken us many thousands of participants to figure exactly which alteration made the difference. (If you’re keeping score, here’s a case where stimulus color *did* matter to the outcome of the experiment!)

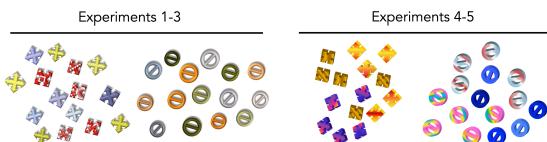


Figure 3.5  
Stimuli from Lewis and Frank (2016) (<https://github.com/mllewis/xtSamp>).

Another explanation for replication failure that is often cited is experimenter expertise (e.g., Schwarz and Strack 2014). On this hypothesis, replications fail because the researchers performing the replication do not have sufficient expertise to execute the study. Like context sensitivity, this explanation is almost certainly true for some replications. In our own work, we have repeatedly performed experiments that failed due to our own incompetence!

Yet as an explanation of the pattern of metascience findings, the expertise hypothesis hasn’t been supported empirically. First, team expertise was not a predictor of replication success in RPP (cf. Bench et al. 2017).

More convincingly, Many Labs 5 selected ten findings from RPP with unsuccessful replications and systematically evaluated whether formal expert peer review of the protocols, including by the authors of the original study, would lead to a larger effect sizes. Despite a massive sample size and extremely thorough review process, there was little to no change in the effects for the vetted protocols relative to the original protocol used

in RPP (Ebersole et al. 2020).

Context, moderators, and expertise seem like reasonable explanations for individual replication failures. Certainly, we should expect them to be explanatory! But for these hypotheses to be operationalized in such a way that they carry weight in our evaluation of the metascientific evidence, they must be evaluated empirically rather than accepted uncritically. When such evaluations have been carried out, they have failed to support a large role for these factors.

1294

1295 The general argument of this chapter is that everything is not all  
1296 right in experimental psychology and, hence, that we need to change  
1297 our methodological practices to avoid negative outcomes like irre-  
1298 producible papers and unreplicable results. Toward that goal, we  
1299 have been presenting metascientific evidence on reproducibility and  
1300 replicability. But this evidence has been controversial, to say the  
1301 least! Do large-scale replication studies like RPP—or for that matter,  
1302 smaller-scale individual replications of effects like “power posing”—  
1303 really lead to the conclusion that our methods require changes? Or are  
1304 there reasons why a lower replication rate is actually consistent with a  
1305 cumulative, positive vision of psychology?

1306 One line of argument addresses this question through the dynamics of  
1307 scientific change. There are many versions, but one is given by Wilson,

1308 Harris, and Wixted (2020). The idea is that progress in psychology con-  
1309 sists of a two-step process by which candidate ideas are “screened” by  
1310 virtue of small, noisy experiments that reveal promising but tentative  
1311 ideas that can then be “confirmed” by large-scale replications. On this  
1312 kind of view, it’s business as usual to find that many randomly selected  
1313 findings don’t hold up in large-scale replications and so we shouldn’t be  
1314 distressed by results like those of RPP. The key to progress is to finding a  
1315 small set that *do* hold up, which will lead to new areas of inquiry. We’re  
1316 not sure that this view is either a good description of current practice  
1317 or a good normative goal for scientific progress, but we won’t focus on  
1318 that critique of Wilson et al.’s argument here. Instead, since this book is  
1319 written for experimenters-in-training, we assume that *you* do not want  
1320 your experiment to be a false positive from a noisy screening procedure,  
1321 regardless of your feelings about the rest of the literature!

1322 In RPP and subsequent metascience studies, original studies with lower  
1323 *p*-values, larger effect sizes, and larger sample sizes were more likely  
1324 to replicate successfully (Yang, Youyou, and Uzzi 2020). From a the-  
1325 oretical perspective, this result is to be expected, because the *p*-value  
1326 literally captures the probability of the data (or any “more extreme”)  
1327 under the null hypothesis of no effect. So a lower *p*-value should indi-  
1328 cate a lower probability of a spurious result.<sup>11</sup> In some sense, the funda-  
1329 mental question about the replication metascience literature is why the

<sup>11</sup> In chapter 6, we will have a lot more to say about  $p < 0.05$ , but for now we’ll mostly just treat it as a particular research outcome.

1330 *p*-values aren't better predictors of replicability! For example, Camerer  
1331 et al. (2018) computes an expected number of successful replications  
1332 on the basis of the effects and sample sizes—and their proportion of suc-  
1333 cessful replications is substantially lower than that number.<sup>12</sup>

1334 One explanation is that the statistical evidence presented in papers often  
1335 dramatically overstates the true evidence from a study. That's because  
1336 of two pervasive and critical issues: **analytic flexibility** (also known as  
1337 *p*-hacking or **questionable research practices**) and **publication bias**.<sup>13</sup>

1338 Publication bias refers to the relative preference (of scientists and other  
1339 stakeholders, like journals) for experiments that "work" over those that  
1340 do not, where "work" is typically defined as yielding a significant result  
1341 at  $p < 0.05$ . Because of this preference, it is typically easier to publish  
1342 positive (statistically significant) results. The relative absence of negative  
1343 results leads to biases in the literature. Intuitively, this bias will lead to a  
1344 literature filled with papers where  $p < 0.05$ . Negative findings will then  
1345 remain unpublished, living in the proverbial "file drawer" (Rosenthal  
1346 1979).<sup>14</sup> In a literature with a high degree of publication bias, many  
1347 findings will be spurious because experimenters got lucky and published  
1348 the study that "worked" even if that success was due to chance variation.  
1349 In this situation, these spurious findings will not be replicable and so the  
1350 overall rate of replicability in the literature will be lowered.<sup>15</sup>

<sup>12</sup> This calculation, as with most other metrics of replication success, assumes that the underlying population effect is exactly the same for the replication and the original. This is a limitation because there could be unmeasured moderators that could produce genuine substantive differences between the two estimates.

<sup>13</sup> These terms basically mean the same thing and are not used very precisely in the literature. *p*-hacking is an informal term that sounds like you know you are doing something bad; sometimes people do, and sometimes they don't. Questionable research practices is a more formal-sounding term that is in principle vague enough to encompass many ethical failings but in practice gets used to talk about *p*-hacking. Unless *p*-hacking intent is crystal clear, we favor two clunkier terms: "data-dependent decision-making" and "undisclosed analytic flexibility." These terms describe the actual practices more precisely: trying many different things after looking at data, typically without reporting all of them.

1351 It's our view that publication bias and its even more pervasive cousin,  
1352 analytic flexibility, are likely to be key drivers of lower replicability. We  
1353 admit that the metascientific evidence for this hypothesis isn't unam-  
1354 biguous, but that's because there's no sure-fire way to diagnose analytic  
1355 flexibility in a particular paper—since we can almost never reconstruct  
1356 the precise choices that were made in the data collection and analysis  
1357 process! On the other hand, it is possible to analyze indicators of publi-  
1358 cation bias in specific literatures, and there are several cases where pub-  
1359 lication bias diagnostics appear to go hand in hand with replication fail-  
1360 ure. For example, in the “power posing” example described above, Sim-  
1361 mons and Simonsohn (2017) noted strong evidence of analytic flexibility  
1362 throughout the literature, leading them to conclude that there was no  
1363 evidential value in the literature. And in the case of “money priming”  
1364 (incidental exposures to images or text about money that were hypoth-  
1365 esized to lead to changes in political attitudes), strong evidence of pub-  
1366 lication bias (Vadillo, Hardwicke, and Shanks 2016) was accompanied  
1367 by a string of failed replications (Rohrer, Pashler, and Harris 2015).

<sup>14</sup> One estimate is that 96% of (not pre-registered) papers report positive findings (Scheel, Schijen, and Lakens 2021). We'll have a lot more to say about analytic flexibility and publication bias in chapters 11 and 16, respectively.

<sup>15</sup> The mathematics of the publication bias scenario strikes some observers as implausible: most psychologists don't run dozens of studies and report only one out of each group (Nelson, Simmons, and Simonsohn 2018). Instead, a more common scenario is to conduct many different analyses and then report the most successful, creating some of the same effects as publication bias—a promotion of spurious variation—without a file drawer full of failed studies.

## ❖ ACCIDENT REPORT

### *Analytic flexibility reveals a fountain of eternal youth*

The way they tell it, Joseph Simmons, Leif Nelson, and Uri Simonsohn wrote their paper on “false positive psychology” (Simmons, Nelson, and

Simonsohn 2011) as an attempt at catharsis (Simmons, Nelson, and Simonsohn 2018). They were fed up with work that they felt exploited flexibility in data analysis to produce findings blessed with  $p < 0.05$  but likely did not reflect replicable effects. They called this practice *p-hacking*: trying different things to get your *p*-value to be below 0.05.

Their paper reported on a simple experiment: they played participants either the Beatles song, “When I’m 64,” or a control song and then asked them to report their date of birth (Simmons, Nelson, and Simonsohn 2011). This manipulation resulted in a significant one and a half year rejuvenation effect. Listening to the Beatles seemed to have made their participants younger!

This result is impossible, of course. But the authors produced a statistically significant difference between the groups that, by definition, was a **false positive**—a case where the statistical test indicated that there was a difference between groups despite no difference existing. In essence, they did so by trying many possible analyses and “cherry-picking” the one that produced a positive result. This practice of course invalidates the inference that the statistical test is supposed to help you make. Several of the practices they followed included:

- selectively reporting dependent measures (e.g., collecting several measures and reporting only one)
- selectively dropping manipulation conditions
- conducting their statistical test and then testing extra participants if

they did not see a significant finding

- adjusting for gender as a covariate in their analysis if doing so resulted in a significant effect

Many of the practices that the authors followed in their rejuvenation study were (and maybe still are!) commonplace in the research literature.

John, Loewenstein, and Prelec (2012) surveyed research psychologists on the prevalence of what they called **questionable research practices**. Most participants admitted to following some of these practices—including exactly the same practices followed by the rejuvenation study.

For many in the field, “false positive psychology” was a galvanizing moment, leading them to recognize how common practices could lead to completely spurious (or even impossible) conclusions. As Simmons, Nelson, and Simonsohn wrote in their article (2018, p. 255), “Everyone knew [ $p$ -hacking] was wrong, but they thought it was wrong the way it is wrong to jaywalk. We decided to write ‘False-Positive Psychology’ when simulations revealed that it was wrong the way it is wrong to rob a bank.”

1370

### <sup>1371</sup> 3.4 Replication, theory building, and open science

<sup>1372</sup> Empirical measures of reproducibility and replicability in the experi-

<sup>1373</sup> mental psychology literature are low—lower than we may have naively

<sup>1374</sup> suspected and lower than we want. How do we address these issues?

<sup>1375</sup> And how do these issues interact with the goal of building theories?

1376 In this section, we discuss the relationship between replication and  
1377 theory—and the role that open and transparent research practices can  
1378 play.

### 1379 *3.4.1 Reciprocity between replication and theory*

1380 Analytic reproducibility is a prerequisite for theory building because if  
1381 the twin goals of theories are to explain and to predict experimental  
1382 measurements, then an error-ridden literature undermines this goal. If  
1383 some proportion of all numerical values reported in the literature were  
1384 simple, unintentional typos, this situation would create an extra level of  
1385 noise—irrelevant random variation—impeding our goal of getting pre-  
1386 cise enough measurements to distinguish between theories. But the  
1387 situation is likely worse: errors are much more often in the direction  
1388 that favors authors' own hypotheses. Thus, irreproducibility not only  
1389 decreases our precision; it also increases the bias in the literature, creat-  
1390 ing obstacles to the fair evaluation of theories with respect to data.

1391 Replicability is also foundational to theory building. Across a range of  
1392 different conceptions of how science works, scientific theories are eval-  
1393 uated with respect to their relationship to the world. They must be  
1394 supported, or at least fail to be falsified, by specific observations. It may

1395 be that some observations are by their nature unrepeatable (e.g., a partic-  
1396 ular astrophysical event might be observed once a human lifetime). But  
1397 for laboratory sciences—and experimental psychology can be counted  
1398 among these, to a certain extent at least—the independent and skeptical  
1399 evaluation of theories requires repeatability of measurements.

1400 Some authors have argued (following the philosopher Heracli-  
1401 tus), “You Cannot Step in the Same River Twice” (McShane and  
1402 Böckenholdt 2014)—meaning that the circumstances and context of  
1403 psychological experiments are constantly changing, and no observation  
1404 will be identical to another. This is of course technically true from a  
1405 philosophical perspective. But that’s where theory comes in! As we  
1406 discussed above, our theories postulate the invariances that allow us to  
1407 group together similar observations and generalize across them.

1408 In this sense, replication is critical to theory, but theory is also critical  
1409 to replication. Without a theory of “what matters” to a particular out-  
1410 come, we really are stepping into an ever-changing river. But a good  
1411 theory can concentrate our expectations on a much smaller set of causal  
1412 relationships, allowing us to make strong predictions about what factors  
1413 should and shouldn’t matter to experimental outcomes. To return to an  
1414 example we discussed earlier, should stimulus color matter to the out-  
1415 come of an experiment? Our theory could tell us that it shouldn’t mat-

<sup>1416</sup> ter for a priming experiment (Baribault et al. 2018) but that it should

<sup>1417</sup> for a generalization experiment (Lewis and Frank 2016).

<sup>1418</sup> *3.4.2 Deciding when to replicate to maximize epistemic value*

<sup>1419</sup> As a scientific community, how much emphasis should we place on

<sup>1420</sup> replication? In the words of Newell (1973), “You Can’t Play 20 Ques-

<sup>1421</sup> tions with Nature and Win.” A series of well-replicated measurements

<sup>1422</sup> does not itself constitute a theory. Theory construction is its own impor-

<sup>1423</sup> tant activity. We’ve tried to make the case here that a reproducible and

<sup>1424</sup> replicable literature is a critical foundation for theory building. That

<sup>1425</sup> doesn’t necessarily mean you have to do replications all the time.

<sup>1426</sup> More generally, any scientific community needs to trade off between

<sup>1427</sup> exploring new phenomena and confirming previously reported effects.

<sup>1428</sup> In a thought-provoking analysis, Oberauer and Lewandowsky (2019)

<sup>1429</sup> suggest that perhaps replications also aren’t the best test of theoretical

<sup>1430</sup> hypotheses. In their analysis, if you don’t have a theory, then it makes

<sup>1431</sup> sense to try and discover new phenomena and then to replicate them.

<sup>1432</sup> If you *do* have a theory, you should expend your energy in testing new

<sup>1433</sup> predictions rather than repeating the same test across multiple replica-

<sup>1434</sup> tions. Analyses such as Oberauer and Lewandowsky (2019) can provide

<sup>1435</sup> a guide to our allocation of scientific effort.

1436 Our goal in this book is somewhat different than the general goal of  
1437 metascientists considering how science should be conducted. Once *you*  
1438 as a researcher decide to do a particular experiment, we think you will  
1439 want to maximize its scientific value and so you will want it to be repli-  
1440 cable. But we aren't suggesting that you should necessarily do a replica-  
1441 tion study. There are many concerns that go into whether to replicate—  
1442 including not only whether you are trying to gather evidence about a  
1443 particular phenomenon but also whether you are trying to master tech-  
1444 niques and paradigms related to it. As we said at the beginning of this  
1445 chapter, not all replication is for the purpose of verification, and you as  
1446 a researcher can make an informed decision about what experimental  
1447 strategy is best for you.

#### 1448 3.4.3 Open science

1449 The **open science movement** is, in part, a response—really a set of  
1450 responses—to the challenges of reproducibility and replicability. The  
1451 open science (and now the broader **open scholarship**) movement is a  
1452 broad umbrella (figure 3.6), but in this book we take open science to be  
1453 a set of beliefs, research practices, results, and policies that are organized  
1454 around the central roles of transparency and verifiability in scientific  
1455 practice.<sup>16</sup> The core of this movement is the idea of “nullius in verba”

<sup>16</sup> Another part of the open science um-  
brella involves a democratization of the  
scientific process through efforts to open  
access to science. This process involves  
both removal of barriers to access the  
scientific literature but also efforts to  
remove barriers to scientific training—  
especially to groups historically under-  
represented in the sciences. The hope  
is that these processes increase both the  
set of people and the range of perspec-  
tives contributing to science. We view  
these changes as no less critical than the  
transparency aspects of the open science  
movement, though more indirectly re-  
lated to the current discussion of repro-  
ducibility and replicability.

1456 (the motto of the British Royal Society, which roughly means “take  
 1457 no one’s word for it”).<sup>17</sup>

1458 Transparency initiatives are critical for ensuring reproducibility. As we  
 1459 discussed above, you cannot even evaluate reproducibility in the ab-  
 1460 sence of data sharing. Code sharing can go even further toward helping  
 1461 reproducibility, as code makes the exact computations involved in data  
 1462 analysis much more explicit than the verbal descriptions that are the  
 1463 norm in papers (Hardwicke et al. 2018). Further, as we will discuss  
 1464 in chapter 13, the set of practices involved in preparing materials for  
 1465 sharing can themselves encourage reproducibility by leading to better  
 1466 organizational practices for research data, materials, and code.

1467 Transparency also plays a major role in advancing replicability. This  
 1468 point may not seem obvious at first—why would sharing things openly  
 1469 lead to more replicable experiments?—but it is one of the major theses  
 1470 of this book, so we’ll unpack it a bit. Here are a couple of routes by  
 1471 which transparent practices lead to greater replication rates.

- 1472 1. Sharing of experimental materials enables replications that closely  
 1473 follow the original study’s methods. One critique of many repli-  
 1474 cations has been that they differ in key respects from the originals.  
 1475 Sometimes those deviations were purposeful, but in other cases

<sup>17</sup> At least that’s a reasonable para-phrase; there’s some interesting discussion about what this quote from Horace really means in a letter by Gould (1991).

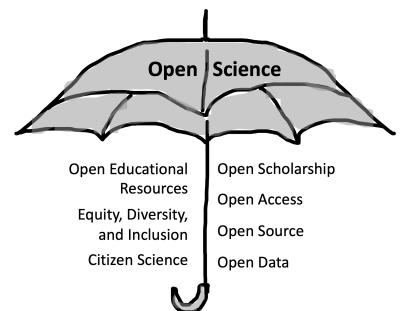


Figure 3.6  
 The broad umbrella of open science (adapted from an image created for the Stanford Lane Library Blog).

1476 they were simply because the replicators could not use the origi-

1477 nal experimental materials. Sharing materials solves this problem.

1478 2. Sharing sampling and analysis plans allows replication of key as-

1479 pects of design and analysis that may not be clear in verbal de-

1480 scriptions, for example exclusion criteria or details of data pre-

1481 processing.

1482 3. Sharing of analytic decision-making via preregistration can lead

1483 to a decrease in *p*-hacking and other practices that can introduce

1484 bias. The strength of statistical evidence in the original study is a

1485 predictor of replicability in subsequent studies. If original studies

1486 are preregistered, they are more likely to report effects that are not

1487 subject to inflation via questionable research practices.

1488 4. Preregistration can also clarify the distinction between confirma-

1489 tory and exploratory findings, helping subsequent experimenters

1490 to make a more informed judgment about which effects are likely

1491 to be good targets for replication.

1492 For all of these reasons, we believe that open science practices can play

1493 a critical role in increasing reproducibility and replicability.

1494 3.4.4 *A crisis?*

1495 So, is there a “replication crisis”? The common meaning of “crisis” is  
1496 “a difficult time.” The data we reviewed in this chapter suggest that  
1497 there are real problems in the reproducibility and replicability of the  
1498 psychology literature. But there’s no evidence that things have gotten  
1499 worse. If anything, we are optimistic about the changes in practices that  
1500 have happened in the last ten years. So in that sense, we are not sure  
1501 that a crisis narrative is warranted.

1502 On the other hand, for Kuhn (1962), the term “crisis” had a special  
1503 meaning: it is a period of intense uncertainty in a scientific field brought  
1504 on by the failure of a particular paradigm (chapter 2). A crisis typically  
1505 heralds a shift in paradigm, in which new approaches and phenomena  
1506 come to the fore.

1507 In this sense, the replication crisis narrative isn’t mutually exclusive with  
1508 other crisis narratives, including the “generalizability crisis” (Yarkoni  
1509 2020) and the “theory crisis” (Oberauer and Lewandowsky 2019). All  
1510 of these are symptoms of discontent with the status quo. We share  
1511 this discontent! We are writing this book to encourage further changes  
1512 in experimental methods and practices to improve reproducibility and  
1513 replicability outcomes—many of them driven by the broader set of ideas  
1514 referred to as “open science.” These changes may not lead to a paradigm

shift in the Kuhnian sense, but we hope that they lead to eventual improvements. In that sense, we think agree with those who say that the “replication crisis” has led to a “credibility revolution” (Vazire 2018).

### 1518 3.5 *Chapter summary: Replication*

1519 In this chapter, we introduce the notions of reproducibility—getting  
1520 the same numbers from the same analysis—and replicability—getting  
1521 the same conclusions from a new dataset. Both of these are critical  
1522 prerequisites of a cumulative scientific literature, yet the metascience  
1523 literature has suggested that the rate of both reproducibility and repli-  
1524 cability in the published literature is quite a bit lower than we would  
1525 hope. A strong candidate explanation for low reproducibility is simply  
1526 that code and data are rarely shared alongside published research. Low-  
1527 ered replicability is more difficult to explain, but our best guess is that  
1528 analytic flexibility (“*p*-hacking”) is at least partially to blame. On our  
1529 account, replication is a metascientific tool for understanding the status  
1530 of the scientific literature rather than an end in itself. Instead, we see  
1531 the open science movement, a movement focused on the role of trans-  
1532 parency in the scientific process, as a promising response to issues of  
1533 reproducibility and replicability.



## DISCUSSION QUESTIONS

1. How would you design a measure of the context sensitivity of an experiment? Think of a measure you could apply post hoc to a description of an experiment (e.g., from reading a paper) so that you could take a group of experiments and annotate how context-sensitive they are on some scale.
2. Take the measure you designed above. How would you test that this measure really captured context sensitivity in a way that was not circular? What would be an “objective measure” of context sensitivity?
3. What proportion of reproducibility failures do you think are due to questionable practices by experimenters vs just plain errors? How would you test your hypothesis?

1534



## READINGS

- Still a very readable and entertaining introduction to the idea of *p*-hacking: Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. (2011). “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11):1359–1366. <https://doi.org/10.1177/0956797611417632>.
- A recent review of issues of replication in psychology: Brian A. Nosek, Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber Almenberg, Fiona Fidler, et al. (2022). “Repli-

1535

cability, Robustness, and Reproducibility in Psychological Science.”

*Annual Review of Psychology* 73 (1): 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>.

1536

## 1537 References

- Anderson, Christopher J, Bahník Štěpán, Barnett-Cowan Michael, Bosco Frank A, Chandler Jesse, Chartier Christopher R, Cheung Felix, et al. 2016. “Response to Comment on “Estimating the Reproducibility of Psychological Science’.” *Science* 351 (6277): 1037.
- Artner, Richard, Thomas Verliefde, Sara Steegen, Sara Gomes, Frits Traets, Francis Tuerlinckx, and Wolf Vanpaemel. 2020. “The Reproducibility of Statistical Results in Psychological Research: An Investigation Using Unpublished Raw Data.” *Psychological Methods* 26 (5): 527–546.
- Bakker, Marjan, and Jelte M. Wicherts. 2011. “The (Mis) Reporting of Statistical Results in Psychology Journals.” *Behavior Research Methods* 43 (3): 666–678.
- Baribault, Beth, Chris Donkin, Daniel R. Little, Jennifer S. Trueblood, Zita Oravecz, Don Van Ravenzwaaij, Corey N. White, Paul De Boeck, and Joachim Vandekerckhove. 2018. “Metastudies for Robust Tests of Theory.” *Proceedings of the National Academy of Sciences* 115 (11): 2607–2612.
- Bench, Shane W., Grace N. Rivera, Rebecca J. Schlegel, Joshua A. Hicks, and Heather C. Lench. 2017. “Does Expertise Matter in Replication? An Examination of the Reproducibility Project: Psychology.” *Journal of Ex-*

1538

- perimental Social Psychology* 68:181–184.
- Buckheit, Jonathan B., and David L. Donoho. 1995. “Wavelab and Reproducible Research.” In *Wavelets and Statistics*, 55–81. Springer.
- Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. “Evaluating Replicability of Laboratory Experiments in Economics.” *Science* 351 (6280): 1433–1436.
- Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. “Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015.” *Nature Human Behaviour* 2 (9): 637–644.
- Carney, Dana R. 2016. “My Position on Power Poses.” *Unpublished Manuscript. Haas School of Business, University of California.* [https://faculty.haas.berkeley.edu/dana\\_carney/pdf\\_my%20position%20on%20power%20poses.pdf](https://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf).
- Carney, Dana R., Amy J. C. Cuddy, and Andy J. Yap. 2010. “Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance.” *Psychological Science* 21 (10): 1363–1368.
- Cesana-Arlotti, Nicoló, Ana Martín, Ernő Téglás, Liza Vorobyova, Ryszard Cetnarski, and Luca L. Bonatti. 2018. “Erratum for the Report ‘Precursors of Logical Reasoning in Preverbal Human Infants’.” *Science* 361 (6408): 1263–1266.
- Dominus, Susan. 2017. “When the Revolution Came for Amy Cuddy.” *New York Times.* <https://www.nytimes.com/2017/10/18/magazine/when-the-revolution-came-foramy-cuddy.html>.

- Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. “Using Prediction Markets to Estimate the Reproducibility of Scientific Research.” *Proceedings of the National Academy of Sciences* 112 (50): 15343–15347.
- Ebersole, Charles R., Maya B. Mathur, Erica Baranski, Diane-Jo Bart-Plange, Nicholas R. Buttrick, Christopher R. Chartier, Katherine S. Corker, et al. 2020. “Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability.” *Advances in Methods and Practices in Psychological Science* 3 (3): 309–331.
- Errington, Timothy M., Maya B. Mathur, Courtney K. Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A. Nosek. 2021. “Investigating the Replicability of Preclinical Cancer Biology.” *Elife* 10:e71601.
- Etz, Alexander, and Joachim Vandekerckhove. 2016. “A Bayesian Perspective on the Reproducibility Project: Psychology.” *PloS One* 11 (2): e0149794.
- Frank, Michael C., and Rebecca Saxe. 2012. “Teaching Replication.” *Perspectives on Psychological Science* 7 (6): 600–604. <https://doi.org/10.1177/1745691612460686>.
- Frank, Michael C., Jonathan A. Slemmer, Gary F. Marcus, and Scott P. Johnson. 2013. “”Information from Multiple Modalities Helps 5-Month-Olds Learn Abstract Rules”: Erratum.” *Developmental Science* 16 (2): 324. <https://doi.org/10.1111/desc.12060>.
- Gelman, Andrew. 2018. “Don’t Characterize Replications as Successes or Failures.” *Behavioral and Brain Sciences* 41.
- Gilbert, Daniel T., Gary King, Stephen Pettigrew, and Timothy D. Wilson. 2016. “Comment on ‘Estimating the Reproducibility of Psychological Sci-

- ence'." *Science* 351 (6277): 1037.
- Gould, Stephen Jay. 1991. "Royal Shorthand." *Science* 251 (4990): 142.
- Gould, Stephen Jay. 1996. *The Mismeasure of Man*. W. W. Norton & Company.
- Hardwicke, Tom E., Manuel Bohn, Kyle MacDonald, Emily Hembacher, Michèle B. Nuijten, Benjamin N. Peloquin, Benjamin E. deMayo, Bria Long, Erica J. Yoon, and Michael C. Frank. 2021. "Analytic Reproducibility in Articles Receiving Open Data Badges at the Journal Psychological Science: An Observational Study." *Royal Society Open Science* 8 (1): 201494. <https://doi.org/10.1098/rsos.201494>.
- Hardwicke, Tom E., Maya B. Mathur, Kyle Earl MacDonald, Gustav Nilsonne, George Christopher Banks, Mallory Kidwell, Alicia Hofelich Mohr, et al. 2018. "Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition." *Royal Society Open Science* 5. <https://doi.org/10.1098/rsos.180448>.
- Hardwicke, Tom E., Robert T. Thibault, Jessica Kosie, Joshua D. Wallach, Mallory C. Kidwell, and John P. A. Ioannidis. 2021. "Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (2014–2017)." *Perspectives on Psychological Science* 17 (1). <https://doi.org/10.1177/1745691620979806>.
- John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23 (5): 524–532.
- Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams Jr, Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs

- 2: Investigating Variation in Replicability across Samples and Settings.” *Advances in Methods and Practices in Psychological Science* 1 (4): 443–490.
- Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Princeton University Press.
- Lewis, Molly L., and Michael C. Frank. 2016. “Understanding the Effect of Social Context on Learning: A Replication of Xu and Tenenbaum (2007b).” *Journal of Experimental Psychology: General* 145 (9): e72.
- Mathur, Maya B., and Tyler J. VanderWeele. 2020. “New Statistical Metrics for Multisite Replication Projects.” *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 183 (3): 1145–1166.
- McShane, Blakeley B., and Ulf Böckenholz. 2014. “You Cannot Step into the Same River Twice: When Power Analyses Are Optimistic.” *Perspectives on Psychological Science* 9 (6): 612–625.
- Nelson, Leif D., Joseph Simmons, and Uri Simonsohn. 2018. “Psychology’s Renaissance.” *Annual Review of Psychology* 69:511–534.
- Newell, Allen. 1973. “You Can’t Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium.” In *Visual Information Processing*, edited by W. G. Chase. Academic Press.
- Nosek, Brian A., Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber Almenberg, Fiona Fidler, Joseph Hilgard, Melissa Kline, and Michèle B. Nuijten. 2022. “Replicability, Robustness, and Reproducibility in Psychological Science.” *Annual Review of Psychology* 73 (1): 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>.
- Nuijten, Michèle B., and Sacha Epskamp. 2024. *statcheck: Extract Statistics from*

*Articles and Recompute p-Values.* <https://CRAN.R-project.org/package=statcheck>.

Nuijten, Michèle B., Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. 2016. “The Prevalence of Statistical Reporting Errors in Psychology (1985–2013).” *Behavior Research Methods* 48 (4): 1205–1226.

Oberauer, Klaus, and Stephan Lewandowsky. 2019. “Addressing the Theory Crisis in Psychology.” *Psychonomic Bulletin & Review* 26 (5): 1596–1618.

Olsson-Collentine, Anton, Jelte M. Wicherts, and Marcel A. L. M. van Assen. 2020. “Heterogeneity in Direct Replications in Psychology and Its Association with Effect Size.” *Psychological Bulletin* 146 (10): 922–940.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251).

Popper, Karl. 1959. *The Logic of Scientific Discovery*. Hutchinson & Co.

Ramscar, Michael. 2016. “Learning and the Replicability of Priming Effects.” *Current Opinion in Psychology* 12:80–84.

Ranehill, Eva, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A. Weber. 2015. “Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women.” *Psychological Science* 26 (5): 653–656.

Rohrer, Doug, Harold Pashler, and Christine R. Harris. 2015. “Do Subtle Reminders of Money Change People’s Political Views?” *Journal of Experimental Psychology: General* 144 (4): e73.

Rosenthal, Robert. 1979. “The File Drawer Problem and Tolerance for Null Results.” *Psychological Bulletin* 86 (3): 638.

Rosenthal, Robert. 1990. “Replication in Behavioral Research.” *Journal of Social Behavior and Personality* 5 (4): 1–30.

Scheel, Anne M., Mitchell R. M. J. Schijen, and Daniël Lakens. 2021. “An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports.” *Advances in Methods and Practices in Psychological Science* 4 (2): 25152459211007467.

Schmidt, Stefan. 2009. “Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences.” *Review of General Psychology* 13:90–100.

Schwarz, Norbert, and Gerald L. Clore. 1983. “Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States.” *Journal of Personality and Social Psychology* 45 (3): 513–523.

Schwarz, Norbert, and Fritz Strack. 2014. “Does Merely Going through the Same Moves Make for a ‘Direct’ Replication? Concepts, Contexts, and Operationalizations.” *Social Psychology* 45 (4): 305–306.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–1366. <https://doi.org/10.1177/0956797611417632>.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2018. “False-Positive Citations.” *Perspectives on Psychological Science* 13 (2): 255–259.

Simmons, Joseph P., and Uri Simonsohn. 2017. “Power Posing: P-Curving the Evidence.” *Psychological Science* 28 (5): 687–693.

Simonsohn, Uri. 2015. “Small Telescopes: Detectability and the Evaluation of Replication Results.” *Psychological Science* 26 (5): 559–569.

- Vadillo, Miguel A., Tom E. Hardwicke, and David R. Shanks. 2016. “Selection Bias, Vote Counting, and Money-Priming Effects: A Comment on Rohrer, Pashler, and Harris (2015) and Vohs (2015).” *Journal of Experimental Psychology: General* 145 (5): 655–663.
- Van Bavel, Jay J., Peter Mende-Siedlecki, William J. Brady, and Diego A. Reinero. 2016. “Contextual Sensitivity in Scientific Reproducibility.” *Proceedings of the National Academy of Sciences* 113 (23): 6454–6459.
- Vazire, Simine. 2018. “Implications of the Credibility Revolution for Productivity, Creativity, and Progress.” *Perspectives on Psychological Science* 13 (4): 411–417.
- Whitaker, K. 2017. “Publishing a Reproducible Paper.” Presentation. <https://doi.org/10.6084/m9.figshare.5440621.v2>.
- Wicherts, Jelte M., Denny Borsboom, Judith Kats, and Dylan Molenaar. 2006. “The Poor Availability of Psychological Research Data for Reanalysis.” *American Psychologist* 61 (7): 726–728.
- Wilson, Brent M., Christine R. Harris, and John T. Wixted. 2020. “Science Is Not a Signal Detection Problem.” *Proceedings of the National Academy of Sciences* 117 (11): 5559–5567.
- Yang, Yang, Wu Youyou, and Brian Uzzi. 2020. “Estimating the Deep Replicability of Scientific Findings Using Human and Artificial Intelligence.” *Proceedings of the National Academy of Sciences* 117 (20): 10762–10768.
- Yarkoni, Tal. 2020. “The Generalizability Crisis.” *Behavioral and Brain Sciences* 45:1–37.
- Youyou, Wu, Yang Yang, and Brian Uzzi. 2023. “A Discipline-Wide Investigation of the Replicability of Psychology Papers over the Past

Two Decades.” *Proceedings of the National Academy of Sciences* 120 (6):

e2208863120.

Zwaan, Rolf Antonius, Alexander Etz, Richard E. Lucas, and Brent Donnel-

lan. 2018. “Making Replication Mainstream.” *Behavioral and Brain Sciences*

 LEARNING GOALS

- Distinguish between consequentialist, deontological, and virtue ethics frameworks
- Identify key ethical issues in performing experimental research
- Discuss ethical responsibilities in analysis and reporting of research
- Describe ethical arguments for open science practices

1548 The fundamental thesis of this book is that experiments are the way  
1549 to estimate causal effects, which are the foundations of theory. And  
1550 as we discussed in chapter 1, the reason why experiments allow for  
1551 strong causal inferences is because of two ingredients: a manipulation—  
1552 in which the experimenter changes the world in some way—and ran-  
1553 domization. Put a different way, experimenters learn about the world  
1554 by randomly deciding to do things to their participants! Is that even  
1555 allowed?<sup>1</sup>

<sup>1</sup> We have placed this chapter in the “Foundations” section of the book because we think it’s critical to start the conversation about your ethical responsibilities as an experimentalist and researcher even before you start planning a study. We’ll come back to the ethical frameworks we describe here in chapter 12, which deals specifically with participant recruitment and the informed consent process.

1557 Experimental research raises a host of ethical issues that deserve consid-  
1558 eration. What can and can't we do to participants in an experiment,  
1559 and what considerations do we owe to them by virtue of their deci-  
1560 sion to participate? To facilitate our discussion of these issues, we start  
1561 by briefly introducing the standard philosophical frameworks for eth-  
1562 ical analysis. We then use those to discuss problems of experimental  
1563 ethics, first from the perspective of participants and then second from  
1564 the perspective of the scientific ecosystem more broadly. We end with  
1565 an ethical argument for TRANSPARENCY.



### CASE STUDY

#### *Shock treatment*

A decade after surviving prisoners were liberated from the last concentration camp, Adolf Eichmann, one of the Holocaust's primary masterminds, was tried for his role in the mass genocide (Baade 1961). While reflecting on his rationale for forcibly removing, torturing, and eventually murdering millions of Jews, an unrepentant Eichmann claimed that he was "merely a cog in the machinery that carried out the directives of the German Reich" and therefore was not directly responsible (Kilham and Mann 1974). This startling admission gave a young researcher an interesting idea: "Could it be that Eichmann and his million accomplices in the Holocaust were just following orders? Could we call them all accomplices?" (Milgram 1974, p. 123).

Stanley Milgram aimed to make a direct test of whether people would comply under the direction of an authority figure no matter how uncomfortable or harmful the outcome. He invited participants into the laboratory to serve as a teacher for an activity (Milgram 1963). Participants were told that they were to administer electric shocks of increasing voltage to another participant, the student, in a nearby room whenever the student provided an incorrect response. In reality, there were no shocks, and the student was an actor who was in on the experiment and only pretended to be in pain when the “shocks” were administered. Participants were encouraged to continue administering shocks despite clearly audible pleas from the student to stop. In one of Milgram’s studies, nearly 65% of participants administered the maximum voltage to the student.

This deeply unsettling result has become, as Ross and Nisbett (1991, p. 55) say, “part of our society’s shared intellectual legacy,” informing our scientific and popular conversation in myriad different ways. At the same time, modern reanalyses of archival materials from the study have called into question whether the deception in the study was effective, casting doubt on its central findings (Perry et al. 2020).

Regardless of its scientific value, Milgram’s study blatantly violates modern ethical norms around the conduct of research. Among other violations, the procedure involved **coercion** that undermined participants’ right to withdraw from the experiment. This coercion appeared to have negative consequences: Milgram noted that a number of his participants displayed anxiety symptoms and nervousness. This observation was dis-

tressing and led to calls for this sort of research to be declared unethical (e.g., Baumrind 1964). The ethical issues surrounding Milgram's study are complex, and some are relatively specific to the particulars of his study and moment (Miller 2009). But the controversy around the study was an important part of convincing the scientific community to adopt stricter policies that protect study participants from unnecessary harm.

1568

#### *1569 4.1 Ethical frameworks*

1570 Was Milgram's experiment (see Case Study) really ethically wrong—in  
1571 the sense that it should not have been performed? You might have the  
1572 intuition that it was unethical due to the harms that the participants  
1573 experienced or the way they were (sometimes) deceived by the experi-  
1574 menter. Others might consider arguments in defense of the experiment,  
1575 perhaps that what we learned from it was sufficiently valuable to justify  
1576 its being conducted. Beyond simply arguing back and forth, how could  
1577 we approach this issue more systematically?

1578 Ethical frameworks offer tools for analyzing such situations. In this sec-  
1579 tion, we'll introduce three of the most commonly used frameworks and  
1580 discuss how each of these could be applied to Milgram's paradigm.

1581 4.1.1 *Consequentialist theories*

1582 Ethical theories provide principles for what constitutes good actions.

1583 The simplest theory of good actions is the **consequentialist theory**: good

1584 actions lead to good results. The most famous consequentialist position

1585 is the **utilitarian position**, originally defined by the philosopher John

1586 Stuart Mill (Flinders 1992). This view emphasizes decision-making

1587 based on the “greatest happiness principle,” or the idea that an action

1588 should be considered morally good based on the degree of happiness

1589 or pleasure people experience because of it, and likewise that an action

1590 should be considered morally bad based on the degree of unhappiness

1591 or pain people experience by the same action (Mill 1859).

1592 A consequentialist analysis of Milgram’s study considers the study’s neg-

1593 ative and positive effects and weighs these against one another. Did the

1594 study cause harm to its participants? On the other hand, did the study

1595 lead to knowledge that prevented harm or caused positive benefits?

1596 Consequentialist analysis can be a straightforward way to justify the risks

1597 and benefits of a particular action, but in the research setting it is unsatis-

1598 fying. Many horrifying experiments would be licensed by a consequen-

1599 tialist analysis and yet feel untenable to us. Imagine that a researcher

1600 forced you to undergo a risky and undesired medical intervention be-

1601 cause the resulting knowledge might benefit thousands of others. This

1602 experiment seems like precisely the kind of thing our ethical framework  
1603 should rule out!

1604 *4.1.2 Deontological approaches*

1605 Harmful research performed against participants' will or without  
1606 their knowledge is repugnant; we consider the Tuskegee Syphilis  
1607 Experiment, a horrifying example of such research (Case Study,  
1608 below). Considering such cases, a few rules seem obvious, for example:  
1609 "Researchers must ask participants' permission before conducting  
1610 research on them." Principles like this one are now formalized in  
1611 all ethical codes for research. They exemplify an approach called  
1612 **deontological** (or duty-based) ethics.

1613 Deontology emphasizes the importance of taking ethically permissible  
1614 actions, regardless of their outcome (Biagetti, Gedutis, and Ma 2020).  
1615 In general, university ethics boards take a deontological approach to  
1616 ethics (Boser 2007). In the context of research, there are four primary  
1617 principles being applied:

1618 1. **Respect for autonomy.** This principle requires that people par-  
1619 ticipating in research studies can make their own decisions about  
1620 their participation, and that those with diminished autonomy

(e.g., children) should receive equal protections (Beauchamp and Childress 2001). Respecting someone's autonomy also means providing them with all the information they need to make an informed decision about whether to participate in a research study (giving **consent**) and giving them further context about the study they have participated in after it is done (**debriefing**).

2. **Beneficence.** This principle means that researchers are obligated to protect the well-being of participants for the duration of the study. Beneficence has two parts. The first is to do no harm. Researchers must take steps to minimize the risks to participants and to disclose any known risks at the onset. If risks are discovered during participation, researchers must notify participants of their discovery and make reasonable efforts to mitigate these risks, even if that means stopping the study altogether. The second is to maximize potential benefits to participants.<sup>2</sup>

3. **Nonmaleficence.** This principle is similar to beneficence (in fact, beneficence and nonmaleficence were a single principle when they were first introduced in the Belmont Report, which we'll discuss later) but differs in its emphasis on doing/causing no harm. In general, harm is bad—but deontology is about intent, not impact, so harm is sometimes warranted when the intent

<sup>2</sup> In practice, this doesn't mean compensating participants with exorbitant amounts of money or gifts, which might cause other issues, like exerting an undue influence on low-income participants to participate. Instead “maximizing benefits” is interpreted as identifying all possible benefits of participation in the research and making them available where possible.

1642       is morally good. For example, administering a vaccine may  
1643       cause some discomfort and pain, but the intent is to protect the  
1644       patient from developing a deadly virus in the future. The harm  
1645       is justifiable under this framework.

1646       **4. Justice.** This principle means that both the benefits and risks of  
1647       a study should be equally distributed among all participants. For  
1648       example, participants should not be systematically assigned to one  
1649       condition over another based on features of their identity such as  
1650       socioeconomic status, race and ethnicity, or gender.

1651       Analyzed from the perspective of these principles, Milgram's study  
1652       raises several red flags. First, Milgram's study reduced participants'  
1653       autonomy by making it difficult for them to voluntarily end their  
1654       involvement (participants were told up to four times to continue  
1655       administering shocks even after they expressed clear opposition).  
1656       Second, the paradigm was designed in a way that it was likely to cause  
1657       harm to its participants by putting them in a very stressful situation.  
1658       Further, Milgram's study may have induced *unnecessary* harm on certain  
1659       participants by failing to screen participants for existing mental health  
1660       issues before beginning the session.

 DEPTH

## *Was Milgram justified?*

Was the harm done in Milgram's experiment justifiable given that it informed our understanding of obedience and conformity? We can't say for sure. What we can say is that in the 10 years following the publication of Milgram's study, the number of papers on (any kind of) obedience increased and the nature of these papers expanded from a focus on religious conformity to a broader interest in social conformity, suggesting that Milgram changed the direction of this research area. Additionally, in a follow-up that Milgram conducted, he reported that 84% of participants in the original study said they were happy to have been involved (Milgram 1974). On the other hand, given concerns about validity in the original study, perhaps its influence on the field was not warranted (Perry et al. 2020).

Many researchers believe there was no ethical way to conduct Milgram's experiment while also protecting the integrity of the research goals, but some have tried. One study recreated a portion of the original experiment, with some critical changes (Burger 2009). Before enrolling in the study, participants completed both a phone screening for mental health concerns, addiction, or extreme trauma, and a formal interview with a licensed clinical psychologist, who identified signs of depression or anxiety. Those who passed these assessments were invited into the lab for a Milgram-type learning study. Experimenters clearly explained that par-

ticipation was voluntary and the decision to participate could be reversed at any point, either by the participant themselves or by a trained clinical psychologist who was present for the duration of the session. Additionally, shock administration never exceeded 150 volts (compared to 450 volts in the original study) and experimenters debriefed participants extensively following the end of the session. This modified replication study found similar patterns of obedience as Milgram's; further, one year later, no participants expressed any indication of stress or trauma associated with their involvement in the study.

1662

1663    4.1.1 *Virtue-based approaches*

1664    A final way that we can approach ethical dilemmas is through a virtue  
1665    framework. A **virtue** is a trait, disposition, or quality that is thought to  
1666    be a moral foundation (Annas 2006). Virtue ethics suggests that people  
1667    can learn to be virtuous by observing those actions in others they admire  
1668    (Morris and Morris 2016). Proponents of virtue ethics say this works for  
1669    two reasons: (1) people are generally good at recognizing morally good  
1670    traits in others and (2) people receive some fulfillment from living virtu-  
1671    ously. Virtue ethics differs from deontology and utilitarianism because  
1672    it focuses on a person's character rather than on the nature of a rule or  
1673    the consequences of an action.

1674    From a research perspective, virtue ethics tells us that in order to behave

1675 virtuously, we must make decisions that consider the context surround-  
1676 ing the experiment (Dillern 2021). In other words, researchers should  
1677 evaluate how their studies might influence a participant's behaviors, es-  
1678 pecially when those behaviors deviate from typical expectations. This  
1679 process is also meant to be adaptive, meaning that researchers must be  
1680 vigilant about both the changing mental states of their participants dur-  
1681 ing the experimental session and whether the planned procedure is no  
1682 longer acceptable.

1683 How can we apply this ethical framework to Milgram's experiment?  
1684 Many virtue ethicists would probably conclude that Milgram's ap-  
1685 proach was neither appropriate (for participants) nor adaptive. Upon  
1686 noticing increasing levels of participant distress, an experimenter  
1687 following the virtue ethics framework should have chosen to end the  
1688 session early or—even better—to have minimized participant distress  
1689 from the beginning.

## 1690 *4.2 Ethical responsibilities to research participants*

1691 Milgram's shock experiment was just one of dozens of unethical hu-  
1692 man subjects studies that garnered the attention and anger of the public  
1693 in the United States. In 1978, the US National Commission for the  
1694 Protection of Human Services of Biomedical and Behavioral Research

1695 released the **Belmont Report**, which described protections for the rights  
1696 of human subjects participating in research studies (Adashi, Walters, and  
1697 Menikoff 2018). Perhaps the most important message found in the re-  
1698 port was the notion that “investigators should not have sole responsibil-  
1699 ity for determining whether research involving human subjects fulfills  
1700 ethical standards. Others, who are independent of the research, must  
1701 share the responsibility.” In other words, ethical research requires both  
1702 transparency and external oversight.

#### 1703 4.2.1 *Institutional review boards*

1704 The creation of **institutional review boards** (IRBs) in the United States  
1705 was an important result of the Belmont Report. While regulatory frame-  
1706 works and standards vary across national boundaries, ethical review of  
1707 research is ubiquitous across countries. In what follows, we focus on  
1708 the US regulatory framework as it has been a model for other ethical  
1709 review systems, but we use the clearer label “ethics review boards” for  
1710 IRBs.

1711 An ethics board is a committee of people who review, evaluate, and  
1712 monitor human subjects research to make sure that participants’ rights  
1713 are protected when they participate in research (Oakes 2002). Ethics  
1714 boards are local; every organization that conducts human subjects or

1715 animal research is required to have its own ethics board or to contract  
1716 with an external one. If you are based at a university, yours likely has  
1717 its own, and its members are probably a mix of scientists, doctors, pro-  
1718 fessors, and community residents.<sup>3</sup>

1719 When a group of researchers have a research question they are interested  
1720 in pursuing with human subjects, they must receive approval from their  
1721 local ethics board before beginning any data collection. The ethics board  
1722 reviews each study to make sure:

1723 1. A study poses no more than **minimal risk** to participants. This  
1724 means the anticipated harm or discomfort to the participant is  
1725 not greater than what would be experienced in everyday life. It is  
1726 possible to perform a study that poses **greater than minimal risk**,  
1727 but it requires additional monitoring to detect any adverse events  
1728 that may occur.

1729 2. Researchers obtain **informed consent** from participants before col-  
1730 lecting any data. This requirement means experimenters must dis-  
1731 close all potential risks and benefits so that participants can make  
1732 an informed decision about whether or not to participate in the  
1733 study. Importantly, informed consent does not stop after partic-  
1734 ipants sign a consent form. If researchers discover any new po-

<sup>3</sup> The local control of ethics boards can lead to very different practices in ethical review across institutions, which is obviously inconsistent with the idea that ethical standards should be uniform! In addition, critics have wondered about the structural issue that institutional ethics boards have an incentive to decrease liability for the institution, while private boards have an incentive to provide approvals to the researchers who pay them (Lemmens and Freedman 2000).

1735 tential risks or benefits along the way, they must disclose these  
1736 discoveries to all participants (see chapter 12).

1737 3. Sensitive information remains **confidential**. Although regulatory  
1738 frameworks vary, researchers typically have an obligation to their  
1739 participants to protect all identifying information recorded during  
1740 the study (see chapter 13).

1741 4. Participants are recruited **equitably** and without **coercion**. Be-  
1742 fore ethics boards became standard, researchers often coercively  
1743 recruited marginalized and vulnerable populations to test their  
1744 research questions, rather than making participation in research  
1745 studies voluntary and providing equitable access to the opportu-  
1746 nity to participate.



### CASE STUDY

#### *The Tuskegee Syphilis Study*

In 1929, The United States Public Health Service (USPHS) was perplexed by the effects of syphilis in Macon County, Alabama, an area with an overwhelmingly Black population (Brandt 1978). Syphilis is a sexually transmitted bacterial infection that can either be in a visible and active stage or in a latent stage. At the time of the study's inception, roughly 36% of Tuskegee's adult population had developed some form of syphilis, one of the highest infection rates in America (White 2006).

The USPHS recruited 400 Black males from 25–60 years of age with latent syphilis and 200 Black males without the infection to serve as a control group to participate (Brandt 1978). The USPHS sought the help of the Macon County Board of Health to recruit participants with the promise that they would provide treatment for community members with syphilis. The researchers sought poor, illiterate Black people and, instead of telling them that they were being recruited for a research study, merely informed them that they would be treated for “bad blood.”

Because the study was interested in tracking the natural course of latent syphilis without any medical intervention, the USPHS had no intention of providing any care to its participants. To assuage participants, the USPHS distributed an ointment that had not been shown to be effective in the treatment of syphilis, and only small doses of a medication actually used to treat the infection. In addition, participants underwent a spinal tap, which was presented to them as another form of therapy and their “last chance for free treatment.”

By 1955, just over 30% of the original participants had died from syphilis complications. It took until the 1970s before the final report was released and (the lack of) treatment ended. In total, 128 participants died of syphilis or complications from the infection, 40 wives became infected, and 19 children were born with the infection (Katz and Warren 2011). The damage rippled through two generations, and many never actually learned what had been done to them.

The Tuskegee experiment violates nearly every single guideline for research described above—indeed in its many horrifying violations of research participants' agency, it provides a blueprint for future regulation to prevent any aspect of it from being repeated: Investigators did not obtain informed consent. Participants were not made aware of all known risks and benefits involved with their participation. Instead, they were deceived by researchers who led them to believe that diagnostic and invasive exams were directly related to their treatment.

Perhaps most shocking, participants were denied appropriate treatment following the discovery that penicillin was effective at treating syphilis (Mahoney, Arnold, and Harris 1943). The USPHS requested that medical professionals overseeing their care outside of the research study not offer treatment to participants so as to preserve the study's methodological integrity. This intervention violated participants' rights to equal access to care, which should have taken precedence over the results of the study.

Finally, recruitment was both imbalanced and coercive. Not only were participants selected from the poorest of neighborhoods in the hopes of finding vulnerable populations with little agency but they were also bribed with empty promises of treatment and a monetary incentive (payment for burial fees, a financial obstacle for many sharecroppers and tenant farmers at the time).

1750 4.2.1 *Risks and benefits*

1751 Imagine that you were approached about participating in a research  
1752 study at your local university. You were only told you would be paid  
1753 \$25 in exchange for completing an hour of cognitive tasks on a com-  
1754 puter. Now imagine that halfway through the session, the experimenter  
1755 revealed they would also need to collect a blood sample, “which should  
1756 only take a couple of minutes and which will really help the research  
1757 study.” Would you agree to the sample? Would you feel uncomfortable  
1758 in any way?

1759 Participants need to understand the risks and benefits of participation in  
1760 an experiment before they give consent. To do otherwise compromises  
1761 their autonomy (a key deontological principle). In the case of this hy-  
1762 pothetical experiment, a new and unexpected invasive component of  
1763 an experiment is coercive: participants would have to choose to forfeit  
1764 their expected compensation to opt out. They also might feel that they  
1765 have been deceived by the experimenter.

1766 In human subjects research, **deception** is a specific technical term that  
1767 refers to cases when (1) experimenters withhold any information about  
1768 its goals or intentions, (2) experimenters hide their true identity (such  
1769 as when using actors), (3) some aspects of the research are under- or  
1770 overstated to conceal information, or (4) participants receive any false

1771 or misleading information. The use of deception requires special con-  
1772 sideration from a human subjects perspective (Kelman 2017; Baumrind  
1773 1985).

1774 Even assuming they are disclosed properly without coercion or decep-  
1775 tion, the risks and benefits of a study must be assessed from the per-  
1776 spective of the *participant*, not the experimenter. By doing so, we allow  
1777 participants to make an informed choice. In the case of the blood sam-  
1778 ple, the risks to the participant were not disclosed and the benefits were  
1779 stated in terms of the research project (and the experimenter).

1780 The benefits of participation in research can either be direct or indi-  
1781 rect, and it is important to specify which type participants may receive.  
1782 While some clinical studies and interventions may offer some direct ben-  
1783 efit due to participation, many of the benefits of basic science research  
1784 are indirect. Both have their place in science, but participants must ul-  
1785 timately determine the degree to which each type of benefit motivates  
1786 their own involvement in a study (Shatz 1986).

1787    4.3 Ethical responsibilities in analysis and reporting of  
1788    research

 ACCIDENT REPORT

*What data?*

Dutch social psychologist Diederick Stapel contributed to more than 200 articles on social comparison, stereotype threat, and discrimination, many published in the most prestigious journals. Stapel reported that affirming positive personal qualities buffered against dangerous social comparison, that product advertisements related to a person's attractiveness changed their sense of self, and that exposure to intelligent in-group members boosted a person's performance on future tasks (Stapel and Linde 2012; Trampe, Stapel, and Siero 2011; Gordijn and Stapel 2006). These findings were fresh and noteworthy at the time of publication, and Stapel's papers were cited thousands of times. The only problem? Stapel's data were made up.

Stapel has admitted that when he first began fabricating data, he would make small tweaks to a few data points (Stapel 2012). Changing a single number here and there would turn a flat study into an impressive one. Having achieved comfortable success (and having aroused little suspicion from journal editors and others in the scientific community), Stapel eventually began creating entire data sets and passing them off as his own. Several colleagues began to grow skeptical of his overwhelming success, however, and brought their concerns to the Psychology Department at

Tilburg University. By the time the investigation of his work concluded, 58 of Stapel's papers were **retracted**, meaning that the publishing journal withdrew the paper after discovering that its contents were invalid.

Everyone agrees that Stapel's behavior was deeply unethical. But should we consider cases of falsification and fraud to be different in kind from other ethical violations in research? Or is fraud merely the endpoint in a continuum that might include other practices like *p*-hacking? Lawyers and philosophers grapple with the precise boundary between sloppiness and neglect, and it can be difficult to know which one is at play when a typo or coding mistake changes the conclusion of a scientific paper. Similarly, if a researcher engages in so-called questionable research practices, at what point should they be considered to have made an ethical violation as opposed to simply performing their research poorly?

The ethical frameworks above provide a framework for thinking about this topic. For the consequentialist, sloppy science can lead to good outcomes for the scientist (quicker publication) but bad outcomes for the rest of the scientific community who have to waste time and effort on papers that may not be correct. For the deontologist, the scientist's intention plays a key role: it is not a generally acceptable principle to knowingly use substandard practices. And for the virtue ethicist, sloppiness is not a morally good trait. On all analyses, researchers have a duty to pursue their work carefully.

1793 also responsible for what we do with our data and for the kinds of con-  
1794 clusions we draw. Cases like Stapel's (see Accident Report) seem stun-  
1795 ning, but they are part of a continuum. Codes of professional ethics for  
1796 organizations like the American Psychological Association encourage  
1797 researchers to take care in the management and analysis of their data so  
1798 as to avoid errors and misstatements (American Psychological Associa-  
1799 tion 2017).

1800 Researchers also have an obligation not to suppress findings based on  
1801 their own beliefs about the right answer. One unfortunate way that  
1802 this suppression can happen is when researchers selectively report their  
1803 research, leading to **publication bias**, as you learned in chapter 3. Re-  
1804 searchers' own biases can be another (invalid) rationale for not publish-  
1805 ing: it's also an ethical violation to suppress findings that contradict your  
1806 theoretical commitments.

1807 Importantly, researchers don't have an obligation to publish *everything*  
1808 they do. Publishing in the peer-reviewed literature is difficult and time-  
1809 consuming. There are plenty of reasons not to publish an experimental  
1810 finding! For example, there's no reason to publish a result if you believe  
1811 it is truly uninformative because of a confound in the experimental de-  
1812 sign. You also aren't typically committing an ethical violation if you de-  
1813 cide to quit your job in research and so you don't publish a study from

1814 your dissertation.<sup>4</sup> The primary ethical issue arises when you use the  
1815 *result* of a study—and how it relates to your own beliefs or to a threshold  
1816 like  $p < 0.05$ —to decide whether to publish it or not.

1817 As we'll discuss again and again in this book, the preparation of research  
1818 reports must also be done with care and attention to detail (see chap-  
1819 ter 14). Sloppiness in writing up results can lead to imprecise or over-  
1820 broad claims; and if that sloppiness extends to the reporting of data, and  
1821 analysis, it may lead to irreproducibility as well.

1822 Further, professional ethics dictates that published contributions to the  
1823 literature be original. In general, the text of a paper must not be pla-  
1824 **giarized** (copied) from the text of other reports whether by you or by  
1825 another author without attribution. Copying from others outside of a  
1826 direct, attributed quotation is obviously an ethical violation because it  
1827 leads to credit for text being given to you rather than the true author.  
1828 But self-plagiarism is also not acceptable—it is a violation to receive  
1829 credit multiple times for the same product.<sup>5</sup>

<sup>4</sup> On the other hand, if your disserta-  
tion contains the cure to a terrible dis-  
ease, you do have a duty to publish it!

<sup>5</sup> Standards on this issue differ from field to field. Our sense is that the rule on self-plagiarism applies primarily to duplication of content between journal papers. So, for example, barring any specific policy of the funder or journal, it is acceptable to use text from one of your own grant proposals in a journal paper. It is also typically acceptable to reuse text from a conference abstract or preregistration (that you wrote, of course) when prepare a journal paper.

1830 4.4 Ethical responsibilities to the broader scientific

1831 community

1832 The open science principles that we will describe throughout this book

1833 are not only important correctives to issues of reproducibility and repli-

1834 cability; they are also ethical duties.

1835 The sociologist Robert Merton described a set of norms that science

1836 is assumed to follow: communism—that scientific knowledge belongs

1837 to the community; universalism—that the validity of scientific results

1838 is independent of the identity of the scientists; disinterestedness—that

1839 scientists and scientific institutions act for the benefit of the overall en-

1840 terprise; and organized skepticism—that scientific findings must be crit-

1841 ically evaluated (Merton 1979).

1842 If the products of science aren't open, it is very hard to be a scientist

1843 by Merton's definition. To contribute to the communal good, papers

1844 need to be openly available. And to be subject to skeptical inquiry, ex-

1845 perimental materials, research data, analytic code, and software must

1846 be all available so that analytic calculations can be verified and experi-

1847 ments can be reproduced. Otherwise, you have to accept arguments on

1848 authority rather than by virtue of the materials and data.

1849 Openness is not only definitionally part of the scientific enterprise; it's

1850 also good for science and individual scientists (Gorgolewski and Pol-  
1851 drack 2016). Publications that are open access are cited more (Eysen-  
1852 bach 2006; Gargouri et al. 2010). Open data also increases the potential  
1853 for citation and reuse, and maximizes the chances that errors are found  
1854 and corrected.

1855 But these benefits mean that researchers have a responsibility to their  
1856 funders to pursue open practices so as to seek the maximal return on  
1857 funders' investments. And by the same logic, if research participants  
1858 contribute their time to scientific projects, the researchers also owe it to  
1859 these participants to maximize the impact of their contributions (Brake-  
1860 wood and Poldrack 2013). For all of these reasons, individual scientists  
1861 have a duty to be open—and scientific institutions have a duty to pro-  
1862 mote transparency in the science they support and publish.

1863 How should these duties be balanced against researchers' other respon-  
1864 sibilities? For example, how should we balance the benefit of data shar-  
1865 ing against the commitment to preserve participant privacy? And, since  
1866 transparency policies also carry costs in terms of time and effort, how  
1867 should researchers consider those costs against other obligations?

1868 First, open practices should be a default in cases where risks and costs  
1869 are limited. For example, the vast majority of journals allow authors to

1870 post accepted manuscripts in their un-typeset form to an open reposi-  
1871 tory. This route to “green” open access is easy, cost free, and—because  
1872 it comes only after articles are accepted for publication—confers essen-  
1873 tially no risks of scooping. As a second example, the vast majority of  
1874 analytic code can be posted as an explicit record of exactly how analy-  
1875 ses were conducted, even if posting data is sometimes more complicated  
1876 due to privacy restrictions. These kinds of “incentive-compatible” ac-  
1877 tions toward openness can bring researchers much of the way to a fully  
1878 transparent workflow, and there is no excuse not to take them.

1879 Second, researchers should plan for sharing and build a workflow that  
1880 decreases the costs of openness. As we discuss in chapter 13, while it  
1881 can be costly and difficult to share data after the fact if they were not  
1882 explicitly prepared for sharing, good project management practices can  
1883 make this process far simpler (and in many cases completely trivial).

1884 Finally, given the ethical imperative toward openness, institutions like  
1885 funders, journals, and societies need to use their role to promote open  
1886 practices and to mitigate potential negatives (Nosek et al. 2015). Schol-  
1887 arly societies have an important role to play in educating scientists about  
1888 the benefits of openness and providing resources to steer their members  
1889 toward best practices for sharing their publication and other research  
1890 products. Similarly, journals can set good defaults, for example by re-

1891 quiring data and code sharing except in cases where a strong justification  
1892 is given. Funders of research can—and increasingly do—signal their in-  
1893 terest in openness through data sharing mandates.

1894 *4.5 Chapter summary: Ethics*

1895 In this chapter, we discussed three ethical frameworks and evaluated  
1896 how they can be applied to our own research through the lens of Mil-  
1897 gram's famous obedience experiment. Studies like Milgram's prompted  
1898 serious conversations about how best to reconcile experimenter goals  
1899 with participant well-being. The publication of the Belmont Report  
1900 and later creation of ethics boards in the United States standardized  
1901 the way scientists approach human subjects research, and created much-  
1902 needed accountability. We also addressed our ethical responsibilities to  
1903 the scientific community, both in how we report our data and how we  
1904 distribute it. We hope that we have convinced you that careful, open  
1905 science is an ethical imperative for researchers!



#### DISCUSSION QUESTIONS

1. The COVID-19 pandemic led to an immense amount of “rapid-response” research in psychology that aimed to discover—and influence—the way people reasoned about contagion, vaccines, masking, and other aspects of the public health situation. What are the

specific ethical concerns that researchers should be aware of for this type of research? Are there reasons for more caution in this kind of research than in other “run-of-the-mill” research?

2. Think of an argument against open science practices—for example, that following open science practices is especially burdensome for researchers with more limited resources (you can make up another if you want!). Given our argument that researchers have an ethical duty to openness, how would you analyze this argument under the three different ethical frameworks we discussed?

1907

## READINGS

- The Belmont Report has shaped US research ethics policy from its publication to the present day. It's also short and quite readable: <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html>.
- A rich reference with several case studies on science misconduct and with strong arguments for open science: Ritchie, Stuart. (2020). *Science fictions: How fraud, bias, negligence, and hype undermine the search for truth*. Metropolitan Books.

1908

1909

||

1910

# STATISTICS

<sup>1911</sup> *References*

- Adashi, Eli Y., LeRoy B. Walters, and Jerry A. Menikoff. 2018. "The Belmont Report at 40: Reckoning with Time." *American Journal of Public Health* 108 (10): 1345–1348.
- American Psychological Association. 2017. "Ethical Principles of Psychologists and Code of Conduct." <https://www.apa.org/ethics/code>.
- Annas, Julia. 2006. "Virtue Ethics." In *The Oxford Handbook of Ethical Theory*, edited by David Copp, 515–536. Oxford University Press.
- Baade, Hans W. 1961. "The Eichmann Trial: Some Legal Aspects." *Duke Law Journal* 10 (3): 400–420.
- Baumrind, Diana. 1964. "Some Thoughts on Ethics of Research: After Reading Milgram's 'Behavioral Study of Obedience'." *American Psychologist* 19 (6): 421.
- Baumrind, Diana. 1985. "Research Using Intentional Deception: Ethical Issues Revisited." *American Psychologist* 40 (2): 165–174.
- Beauchamp, Tom L., and James F. Childress. 2001. *Principles of Biomedical Ethics*. Oxford University Press.
- Biagetti, Maria Teresa, Aldis Gedutis, and Lai Ma. 2020. "Ethical Theories in Research Evaluation: An Exploratory Approach." *Scholarly Assessment Reports* 2 (1): 1–9.
- Boser, Susan. 2007. "Power, Ethics, and the IRB: Dissonance over Human Participant Review of Participatory Research." *Qualitative Inquiry* 13 (8): 1060–1074.
- Brakewood, Beth, and Russell A. Poldrack. 2013. "The Ethics of Secondary Data Analysis: Considering the Application of Belmont Principles to the

- Sharing of Neuroimaging Data.” *Neuroimage* 82:671–676.
- Brandt, Allan M. 1978. “Racism and Research: The Case of the Tuskegee Syphilis Study.” *Hastings Center Report* 8 (6): 21–29.
- Burger, Jerry M. 2009. “Replicating Milgram: Would People Still Obey Today?” *American Psychologist* 64 (1): 1–11.
- Dillern, Thomas. 2021. “The Scientific Judgment-Making Process from a Virtue Ethics Perspective.” *Journal of Academic Ethics* 19 (4): 501–516.
- Eysenbach, Gunther. 2006. “Citation Advantage of Open Access Articles.” *PLoS Biology* 4 (5): e157.
- Flinders, David J. 1992. “In Search of Ethical Guidance: Constructing a Basis for Dialogue.” *International Journal of Qualitative Studies in Education* 5 (2): 101–115.
- Gargouri, Yassine, Chawki Hajjem, Vincent Larivière, Yves Gingras, Les Carr, Tim Brody, and Stevan Harnad. 2010. “Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research.” *PLoS One* 5 (10): e13636.
- Gordijn, Ernestine H., and Diederik A. Stapel. 2006. “Retracted: Behavioural Effects of Automatic Interpersonal Versus Intergroup Social Comparison.” *British Journal of Social Psychology* 45 (4): 717–729. <https://doi.org/10.1348/014466605X79589>.
- Gorgolewski, Krzysztof J., and Russell A. Poldrack. 2016. “A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research.” *PLoS Biology* 14 (7): e1002506. <https://doi.org/10.1371/journal.pbio.1002506>.
- <sup>1913</sup> Katz, Ralph V., and Rueben C. Warren. 2011. *The Search for the Legacy of the*

- USPHS Syphilis Study at Tuskegee.* Lexington Books.
- Kelman, Herbert C. 2017. “Human Use of Human Subjects: The Problem of Deception in Social Psychological Experiments.” In *Research Design*, 189–204. Routledge.
- Kilham, Wesley, and Leon Mann. 1974. “Level of Destructive Obedience as a Function of Transmitter and Executant Roles in the Milgram Obedience Paradigm.” *Journal of Personality and Social Psychology* 29 (5): 696.
- Lemmens, Trudo, and Benjamin Freedman. 2000. “Ethics Review for Sale? Conflict of Interest and Commercial Research Review Boards.” *The Milbank Quarterly* 78 (4): 547–584.
- Mahoney, John F., R. C. Arnold, and Ad Harris. 1943. “Penicillin Treatment of Early Syphilis—a Preliminary Report.” *American Journal of Public Health and the Nations Health* 33 (12): 1387–1391.
- Merton, Robert K. 1979. “The Normative Structure of Science.” *The Sociology of Science: Theoretical and Empirical Investigations*, 267–278.
- Milgram, Stanley. 1963. “Behavioral Study of Obedience.” *The Journal of Abnormal and Social Psychology* 67 (4): 371.
- Milgram, Stanley. 1974. *Obedience to Authority: An Experimental View*. Harper & Row.
- Mill, John Stuart. 1859. “Utilitarianism (1863).” *Utilitarianism, Liberty, Representative Government*, 7–9.
- Miller, Arthur G. 2009. “Reflections on ‘Replicating Milgram’ (Burger, 2009).” *American Psychologist* 64 (1): 20–27. <https://doi.org/10.1037/a0014407>.
- <sup>1914</sup> Morris, Marilyn C., and Jason Z. Morris. 2016. “The Importance of Virtue

- Ethics in the IRB.” *Research Ethics* 12 (4): 201–216.
- Nosek, Brian A., George Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck, et al. 2015. “Promoting an Open Research Culture.” *Science* 348 (6242): 1422–1425. <https://doi.org/10.1126/science.aab2374>.
- Oakes, J. Michael. 2002. “Risks and Wrongs in Social Science Research: An Evaluator’s Guide to the IRB.” *Evaluation Review* 26 (5): 443–479.
- Perry, Gina, Augustine Brannigan, Richard A. Wanner, and Henderikus Stam. 2020. “Credibility and Incredulity in Milgram’s Obedience Experiments: A Reanalysis of an Unpublished Test.” *Social Psychology Quarterly* 83 (1): 88–106.
- Ritchie, Stuart. 2020. *Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth*. Metropolitan Books.
- Ross, Lee, and Richard E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. McGraw-Hill.
- Shatz, David. 1986. “Autonomy, Beneficence, and Informed Consent: Re-thinking the Connections.” *Cancer Investigation* 4 (3): 257–269.
- Stapel, Diederik A. 2012. *Ontsporing*. Prometheus Amsterdam.
- Stapel, Diederik A., and Lonneke A. J. G. van der Linde. 2012. ““What Drives Self-Affirmation Effects? On the Importance of Differentiating Value Affirmation and Attribute Affirmation’: Retraction of Stapel and van Der Linde (2011).” *Journal of Personality and Social Psychology* 103 (3): 505. <https://doi.org/10.1037/a0029745>.
- Trampe, Debra, Diederik A. Stapel, and Frans W. Siero. 2011. “Retracted: The Self-Activation Effect of Advertisements: Ads Can Affect Whether

and How Consumers Think about the Self.” *Journal of Consumer Research*

37 (6): 1030–1045.

White, Robert M. 2006. “Effects of Untreated Syphilis in the Negro Male,

1932 to 1972: A Closure Comes to the Tuskegee Study, 2004.” *Urology* 67

1916 (3): 654.

## <sup>1917</sup> 5 ESTIMATION

### LEARNING GOALS

- Estimate the causal effect of a manipulation
- Discuss differences between frequentist and Bayesian estimation
- Reason about standardized effect sizes and their strengths and weaknesses
- Quantify the relationship between variables

<sup>1918</sup>

<sup>1919</sup> In every quantitative paper we read, every quantitative  
<sup>1920</sup> talk we attend, and every quantitative article we write,  
<sup>1921</sup> we should all ask one question: *what is the estimand?* The  
<sup>1922</sup> estimand is the object of inquiry—it is the precise quantity  
<sup>1923</sup> about which we marshal data to draw an inference. Yet,  
<sup>1924</sup> too often social scientists skip the step of defining the  
<sup>1925</sup> estimand. Instead, they leap straight to describing the  
<sup>1926</sup> data they analyze and the statistical procedures they apply.

1927 Without a statement of the estimand, it becomes impossi-  
1928 ble for the reader to know whether those procedures were  
1929 appropriate.

1930 —Lundberg, Johnson, and Stewart (2021, p. 532)

1931 In the first section of this book, our goal was to set up some of the the-  
1932 oretical ideas that motivate our approach to experimental design and  
1933 planning. We introduced our key thesis, namely that experiments are  
1934 about measuring causal effects. We also began to discuss our key themes:

1935 TRANSPARENCY, MEASUREMENT PRECISION, BIAS REDUCTION, and GENER-  
1936 ALIZABILITY.

1937 In this next section of the book—treating statistical topics—we will inte-  
1938 grate these ideas with an analytic toolkit for estimating effects and quan-  
1939 tifying their size (this chapter), making inferences about how these esti-  
1940 mates relate to a population (chapter 6), and building models for estima-  
1941 tion and inference in more complex settings (chapter 7). Although this  
1942 book does not provide an extensive treatment of statistics, we hope that  
1943 these chapters provide a foundation—and an opinionated perspective—  
1944 for beginning the statistical analysis of your experimental data, with a  
1945 focus on MEASUREMENT PRECISION.



## CASE STUDY

*The Lady Tasting Tea*

The birth of modern statistical inference arose from the age-old conundrum of how to best make a cup of tea. The statistician Ronald Fisher was apparently at an afternoon tea party when a lady declared that she could tell the difference when tea was added to milk vs milk to tea. Rather than taking her at her word, Fisher devised an experimental and data analysis procedure to test her claim.

The lady would have to judge a set of six new cups of tea and sort them into milk-first vs tea-first sets. Her data would then be analyzed to determine whether her level of correct choice exceeded that expected by chance. While this process now sounds like a quotidian experiment that might be done on a cooking reality show, it seems unremarkable in hindsight only because it set the standard for the way science was done going forward.

The important and unusual element of the experiment was its treatment of potential design confounds such which cup of tea was prepared first, which cup of tea was presented first, or the material that the cups were made out of. Prior experimental practice would have been to try to equate all of the cups as closely as possible, decreasing the influence of confounds. Fisher recognized that this strategy was insufficient because of the presence of unobserved confounders. Only by randomizing all other aspects of the experiment could he make strong causal inferences about the

treatment (milk then tea vs tea then milk). We discussed the causal power of random assignment in chapter 1—the Lady Tasting Tea experiment is a key touchstone in the popularization of randomized experiments!

1947

## 1948 5.1 Estimating a quantity

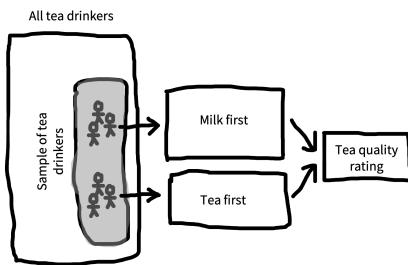


Figure 5.1  
The structure of our tea-tasting experiment.

1949 If experiments are about estimating effects, how do we actually use our  
 1950 experimental data to make these estimates? For our example we'll de-  
 1951 sign a slightly more modern version of Fisher's experiment, shown in  
 1952 figure 5.1.

1953 Our causal theory is that the tea quality is affected by milk-tea ordering,  
 1954 so we'll test that by rating tea quality both milk-first and tea-first, rep-  
 1955 resented by a DAG like the one in figure 5.2. Our intended population  
 1956 to generalize to is the set of all tea drinkers, and toward that goal we  
 1957 sample a set of tea drinkers. In practice, we might do a field trial in a  
 1958 cafe in which we approach patrons and ask them to participate in our  
 1959 experiment in exchange for a free cup of tea. Although this sample size

An important piece of context for the work of Ronald Fisher, Karl Pearson, and other early pioneers of statistical inference is that they were all strong proponents of eugenics. Fisher was the founding chairman of the Cambridge Eugenics Society. Pearson was perhaps even worse, an avowed Social Darwinist who believed fervently in eugenic legislation. These views are repugnant and provide important context for their statistical contributions.

1960 is almost certainly too small to get precise estimates, for the purpose of  
 1961 this example, we'll sample 18 tea drinkers—nine in each condition.

1962 As our manipulation, we follow Fisher in randomly assigning partici-  
 1963 pants (who of course should give consent to participate) into to one of  
 1964 our two conditions: milk-first and tea-first.<sup>1</sup> This design is a between-  
 1965 participants design, so each participant gets only one cup of tea. They  
 1966 receive their cup of tea and taste it. Then, as our measure, we ask for  
 1967 a rating of the tea on a continuous scale from 1 (terrible) to 7 (deli-  
 1968 cious).<sup>2</sup>

1969 An example dataset from our experiment is shown in figure 5.3. Eventu-  
 1970 ally, we'll want to estimate the effect of milk-first preparation on quality  
 1971 ratings (our effect of interest). But for now, our goal will be to estimate  
 1972 the quality of the tea when it is milk-first (which some data suggest is  
 1973 actually the better way, at least for British tea drinkers; Kennedy 2003).  
 1974 More formally, we want to use our sample of nine milk-first tea judg-  
 1975 ments to estimate a number that we can't directly observe, namely the  
 1976 true perceived quality of all possible milk-first cups. We'll call this num-  
 1977 ber a **population parameter** for reasons that will become clear in a mo-  
 1978 ment.

1979 We'll try to go easy on notation, but some amount will hopefully make  
 1980 things clearer. We will use  $\theta_M$  ("theta") to denote the parameter we

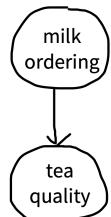


Figure 5.2  
 A directed acyclic graph representing our causal theory of tea quality.

<sup>1</sup> Technically, randomized experiments were not invented by Fisher. Perhaps the earliest example of a (somewhat) randomized experiment was a trial of scurvy treatments in the 1700s (Dunn 1997). Peirce and Jastrow (1884) also report a strikingly modern use of randomized stimulus presentation (via shuffling cards). Nevertheless, Fisher's statistical work popularized randomized experiments throughout the sciences, in part by integrating them with a set of analytic methods.

1981 want to estimate (the population parameter) and  $\hat{\theta}_M$ , its sample esti-  
 1982 mate.<sup>3</sup>

### 1983 5.1.1 Maximum likelihood estimation

1984 Okay, you are probably saying, if we want our estimate of milk-first  
 1985 quality, shouldn't we just take the average rating across the nine cups  
 1986 of milk-first tea? The answer is yes. But let's unpack that choice: tak-  
 1987 ing the sample mean as our estimate  $\hat{\theta}_M$  is an example of an estimation  
 1988 approach called **maximum likelihood estimation**. In general terms, max-  
 1989 imum likelihood estimation is a two-step process.

1990 First, we assume a **model** for how the data were generated.<sup>4</sup> This model  
 1991 is specified in terms of certain population parameters. In our example,  
 1992 the model is as simple as they come: we just assume there is some aver-  
 1993 age level of tea quality and that the measurements vary around it. Let's  
 1994 take a look at the data from the milk-first condition, shown in figure 5.4.  
 1995 Our observations are clustered around the mean, but they also show  
 1996 some variation. Some are higher and some are lower. Variation of this  
 1997 type is a feature of every data set. This variation can be summarized  
 1998 via a **probability distribution**, a mathematical entity that describes the  
 1999 properties of possible datasets.

<sup>2</sup> Right now we're going to assume that our ratings are just simple numerical values and not worry about the fact that they come from a rating scale that is bounded (e.g., can't go above 7). If you're curious about **Likert scales** (the name for discrete numerical rating scales; pronounced LICK-ERT), we'll talk a bit more about them in chapter 8.

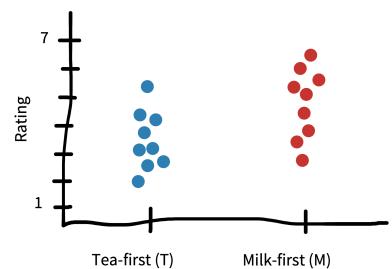


Figure 5.3  
Schematic data from the tea-tasting experiment.

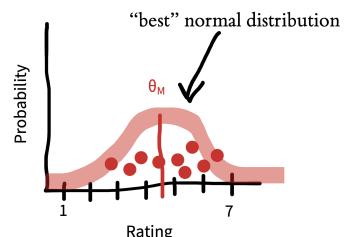
<sup>3</sup> Statisticians use "hats" like this to denote estimates from a specific sample. One way to remember this is that the "person in the hat" is wearing a hat to dress up as the actual quantity.

2000 The only probability distribution we'll discuss here is the ubiquitous  
 2001 **normal distribution** (also sometimes called a "Gaussian distribution"). A  
 2002 normal distribution has two **parameters** (numbers that define its shape):  
 2003 a **mean** and a **standard deviation**. These two parameters define the shape  
 2004 of the curve. The mean ( $\theta_M$ ) describes where its center goes, and the  
 2005 standard deviation describes how wide it is. Technically, the mean is  
 2006 the **expected value** for new samples from the distribution. Our best  
 2007 guess about the value of these new samples is that they are at the mean.  
 2008 We can write this more formally by introducing  $E[M]$  to denote the  
 2009 expectation of the variable  $M$ .

2010 The standard deviation  $\sigma_M$  is then a way of describing the expected  
 2011 *variation* in these samples. A bigger standard deviation means that we  
 2012 expect samples to be on average further from the mean. We can write  
 2013 this formally as  $\sigma_M = \sqrt{E[(M - \theta_M)^2]}$ : the standard deviation is the  
 2014 expected absolute distance between individual samples and the mean,  
 2015 with the square and square root being necessary to compute distance.

2016 Using a probability distribution to describe our dataset gives us a way of  
 2017 summarizing our observations through the parameters of the distribu-  
 2018 tion and encoding an assumption about what future observations might  
 2019 look like. How do we fit a normal distribution to our data? We try  
 2020 to find the values of the population parameters that make our observed

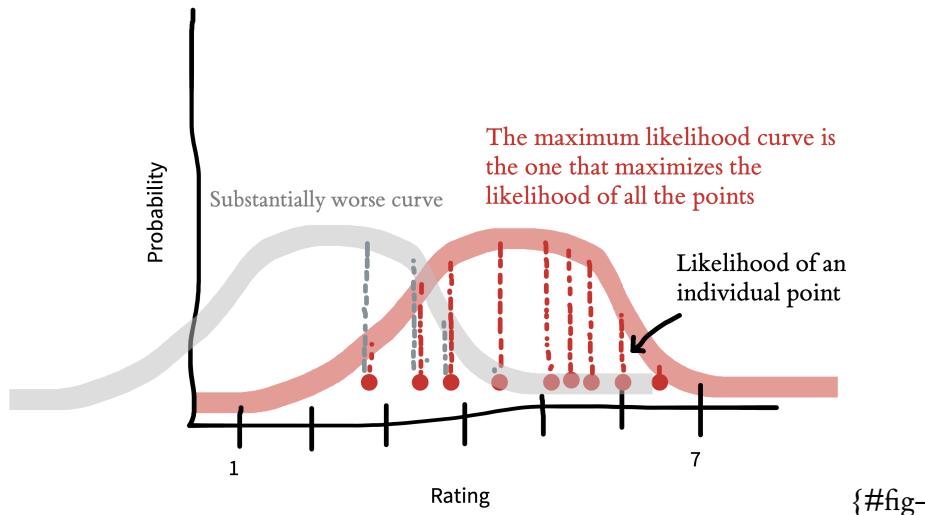
<sup>4</sup> This sense of "model" is actually a formal instantiation of the type of causal model we discussed in chapter 1. As you get deeper into causal modeling, typically what you do is define a causal "story" for the statistical process that generated a dataset, using both DAGs and the kinds of probability distributions we define below.



$\theta_M$  is the mean of the best-fitting normal distribution

Figure 5.4  
The best-fitting normal distribution for data from the milk-first condition.

2021 data as likely as possible. Let's start with the mean.



2022 {#fig-  
 2023 estimation-ml2 .margin-caption width=60% fig-alt="A plot with  
 2024 a "substantially worse curve" and "maximum likelihood curve", each  
 2025 point has dashed line to curve."}  
 2026 }

2026 For example, if our sample mean is  $\hat{\theta}_M = 4.5$ , what underlying value  
 2027 of  $\theta_M$  would make these data most likely to occur? Well, suppose the  
 2028 underlying parameter were  $\theta_M = 2.5$ . Then it would be pretty unlikely  
 2029 that our sample mean would be so much bigger. So,  $\hat{\theta}_M = 2.5$  is a  
 2030 poor estimate of the population parameter based on these data (?@fig-  
 2031 estimation-ml2). Conversely, if the parameter were  $\theta_M = 6.5$ , it would  
 2032 be a bit unlikely that our sample mean would be so much *smaller*. The  
 2033 value of  $\hat{\theta}_M$  that makes these data most likely is just 4.5 itself: the sam-  
 2034 ple mean! That is why the sample mean in this case is the maximum  
 2035 likelihood estimate.

2036 5.1.2 Bayesian estimation

2037 The maximum likelihood estimation example above describes a  
2038 common approach to estimating parameters, where the researcher  
2039 completely puts aside their prior expectations about what these values  
2040 might be. This approach is an example of a **frequentist** statistical  
2041 approach, an approach that focuses on the long-run performance of  
2042 estimation procedures.

2043 Often this approach makes sense, especially when we have no prior ex-  
2044 pectations about the values we are estimating. But sometimes we *do*  
2045 have relevant beliefs about the value. For example, before we perform  
2046 our tea experiment, we don't know exactly what  $\theta_M$  will be, but it seems  
2047 a bit unlikely that tea would be consistently rated as either horrible (1)  
2048 or perfect (7). We have what you might call *weak prior expectations* about  
2049 the kinds of ratings we'll receive.

2050 These kind of expectations are most useful when we have a very small  
2051 amount of data. Remember that our goal is to estimate a population  
2052 parameter using the sample data, and small data sets can be rather noisy.  
2053 Taking into account our prior expectations can help to temper the in-  
2054 fluence of noise. For example, if our very first participant in the ex-  
2055 periment rated their tea as terrible, we wouldn't want to jump to the  
2056 conclusion that the tea was actually bad. Instead, we might speculate

2057 that the participant was having a bad day or had just brushed their teeth.  
 2058 On the other hand, if all of our participants gave bad ratings to their tea,  
 2059 the data would be more persuasive; in that case, we might want to tell  
 2060 the cafe that they are serving substandard tea. The extent to which our  
 2061 prior expectations should moderate our conclusions should vary with  
 2062 the amount of sample data; with only a little data, our prior expecta-  
 2063 tions should have more influence, but as we gather more, we should  
 2064 put greater weight on the data.

2065 How do we quantify this trade-off between our prior expectations and  
 2066 our current observations? We can do this via **Bayesian estimation** of  
 2067  $\hat{\theta}_M$ . Bayesian estimation provides a principled framework for integrating  
 2068 prior beliefs and data. These estimation techniques can be very helpful  
 2069 in cases where data are sparse or prior beliefs are strong.

2070 In Bayesian estimation, we observe some data  $d$ , consisting of the set  
 2071 of responses in the experiment. Now we can use **Bayes's rule**, a tool  
 2072 from basic probability theory, to estimate this number (figure 5.5). Each  
 2073 part of this equation has a name, and it's worth becoming familiar with  
 2074 them. The thing we want to compute,  $p(\theta_M|data)$ , is called the **posterior**  
 2075 **probability**—it tells us what we should believe about the population  
 2076 parameter on tea quality, given the data we observed.<sup>5</sup>

2077 The first part of the numerator is  $p(data|\theta_M)$ , the probability of the data

The diagram illustrates Bayes's rule with the following components:

- posterior**: A purple arrow points to the term  $p(\theta_M|data)$ .
- likelihood**: A red arrow points to the term  $p(data|\theta_M)$ .
- prior**: A blue arrow points to the term  $p(\theta_M)$ .
- The formula is shown as: 
$$p(\theta_M|data) = \frac{p(data|\theta_M) p(\theta_M)}{p(data)}$$

Figure 5.5  
Bayes's rule, annotated.

<sup>5</sup> We're making the posterior **purple** to indicate the combination of likelihood (**red**) and prior (**blue**).

2078 we observed given our hypothesis about the participant’s ability. This  
 2079 part is called the **likelihood**.<sup>6</sup> This term tells us about the relationship  
 2080 between our hypothesis and the data we observed—so, if we think the  
 2081 tea is of high quality (say  $\theta_M = 6.5$ ), then the probability of observing a  
 2082 bunch of low-quality ratings will be fairly low.

2083 The second term in the numerator,  $p(\theta_M)$ , is called the **prior**. This term  
 2084 encodes our beliefs about the likely distribution of tea quality. Intu-  
 2085 itively, if we think that the tea is likely of high quality, we should re-  
 2086 quire more evidence to convince us that it’s bad. In contrast, if we think  
 2087 it’s probably bad, a few examples of low ratings might serve to convince  
 2088 us.

2089 Figure 5.6 gives an example of the combination of prior and data. In  
 2090 this example, we look at what difference the prior makes after observ-  
 2091 ing 9 ratings. If we go in assuming that the tea is likely to be bad, the  
 2092 posterior mean (purple line) will be pushed downward relative to the  
 2093 maximum likelihood estimate (red line). This prior is operating only  
 2094 over on ratings—estimates of tea quality. Later on when we talk about  
 2095 comparing milk-first and tea-first ratings to get an estimate of the ex-  
 2096 perimental effect, we could consider putting a prior on tea *discrimination*  
 2097 (e.g., the experimental effect).

6 Speaking informally, “likelihood” is just a synonym for probability, but in Bayesian estimation, “likelihood” is a technical term specifically referring to probability of the data given our hypothesis. This ambiguity can get a bit confusing.

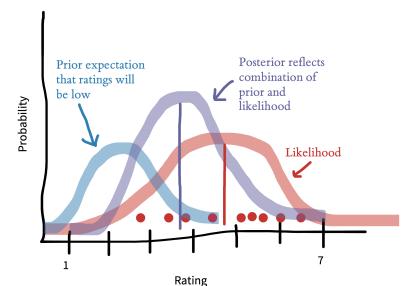


Figure 5.6  
Bayesian inference about tea ratings with a strong prior on low values.

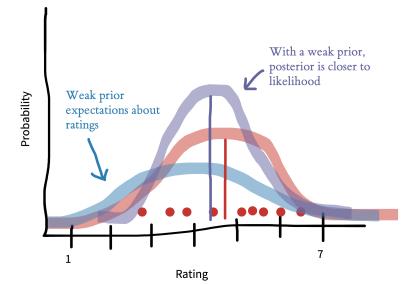


Figure 5.7  
Bayesian inference about tea ratings with a weak prior on low values.

2098 Priors aren’t usually as strong as the one shown above. Figure 5.7 shows

2099 how the picture shifts when we have a weaker prior reflecting a flatter,  
 2100 more widely spread belief about the distribution of ratings. Now the  
 2101 posterior mean (purple) is closer to the maximum likelihood mean (red).  
 2102 This situation is more common—the prior encodes a weak assumption  
 2103 that ratings won't cluster around the ends of the scale.

2104 The effect of the prior is also decreased when you have more data. Take  
 2105 a look at figure 5.8. The prior is the same as in figure 5.6, but we  
 2106 have more data. As a result, the posterior distribution is much more  
 2107 peaked and also much closer to the data—the prior makes much less  
 2108 difference.

2109 Bayesian estimation is most important when you have strong beliefs  
 2110 and not a lot of data. That can be a case where you have just a few  
 2111 participants in your experiment, but it's also good—and perhaps more  
 2112 common—to use Bayesian methods when you have a lot of data, but  
 2113 maybe not that much data about particular units that you care about.  
 2114 For example, you might have a large dataset about the effects of an ed-  
 2115 uational intervention but not that much data about how it affects a  
 2116 particular subgroup. Bayesian estimates and maximum likelihood esti-  
 2117 mates will exactly coincide either under a flat prior (a prior that makes  
 2118 any value equally likely) or as the amount of data goes to infinity.

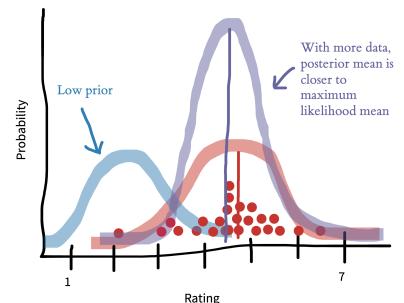


Figure 5.8  
 Bayesian inference about tea ratings with a strong prior on low values and more data.

## 2119 5.2 Estimating and comparing effects

2120 We've now covered estimating a single parameter (the mean for people  
 2121 who had milk-first tea) using both frequentist and Bayesian methods.  
 2122 But recall that what we really wanted to do was to estimate the *causal*  
 2123 *effect* we were interested in, namely the milk-first vs tea-first effect. In  
 2124 this section, we'll discuss how to estimate the effect, and then how to  
 2125 use **effect size** measures to compare effects across experiments (as well  
 2126 as some of the pros and cons of doing so).<sup>7</sup>

### 2127 5.2.1 Estimating the treatment effect

2128 Let's refer to the causal effect we care about as our **treatment effect**.<sup>8</sup>  
 2129 In practice, estimating  $\beta$  (a parameter describing the treatment effect) is  
 2130 going to be a pretty straightforward extension to what we did before.

2131 In the maximum likelihood framework, we could posit that ratings in  
 2132 each group (milk-first and tea-first) follow a normal distribution but  
 2133 that these normal distributions might have different means and standard  
 2134 deviations. Extending the notation introduced above, let's term the pa-  
 2135 rameters for the tea-first group  $\theta_T$  (the mean) and  $\sigma$  (the standard devia-  
 2136 tion). To estimate the treatment effect, we are positing a **model** in which  
 2137 the milk-first ratings are normally distributed with mean  $\theta_M = \theta_T + \beta$   
 2138 and with standard deviation  $\sigma$ .<sup>9</sup> This equation says that milk-first ratings

<sup>7</sup> This method doesn't have to be used only with a causal effect; it can be any between-group difference. In the current example, we can say with certainty that this effect is a causal one because our experiment uses random assignment.

<sup>8</sup> This is the effect of our manipulation—what we sometimes call an “intervention” as well. “Treatment” is a term that comes from medical statistics but is used more broadly in statistics now.

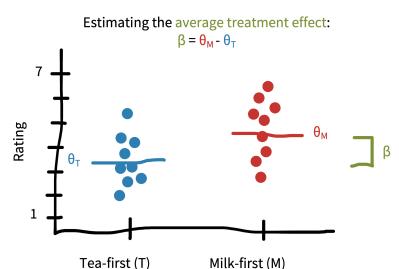


Figure 5.9  
 Estimating the average treatment effect from the tea-tasting data.

<sup>9</sup> For simplicity, we're assuming that the standard deviations in each tea group are equal.

<sup>2139</sup> have the same distribution as tea-first ratings, except that their average  
<sup>2140</sup> is shifted by  $\beta$ . Setting our model up this way then lets us compute  $\hat{\beta}$ ,  
<sup>2141</sup> our estimate of the treatment effect in our sample.

<sup>2142</sup> As in the one-sample case (i.e., estimating the mean of just the milk-  
<sup>2143</sup> first group), maximum likelihood estimation would then proceed by  
<sup>2144</sup> finding the value of  $\beta$  that makes the data most likely under the assumed  
<sup>2145</sup> model. As you'd probably expect, this estimate  $\hat{\beta}$  turns out to be simply  
<sup>2146</sup> the difference in sample means,  $\hat{\theta}_M - \hat{\theta}_T$ . You can see this difference  
<sup>2147</sup> pictured in figure 5.9.

<sup>2148</sup> In the Bayesian framework, we would again specify a prior  $p(\beta)$  that  
<sup>2149</sup> encodes our prior beliefs about the size and direction of the treatment  
<sup>2150</sup> effect. If we have no prior beliefs at all, then we could specify a flat  
<sup>2151</sup> prior,  $p(\beta) \propto 1$ .<sup>10</sup> If we believe the treatment effect is likely to favor  
<sup>2152</sup> milk-first pouring ( $\beta > 0$ ), we could specify that the prior is a normal  
<sup>2153</sup> distribution centered at some positive value (e.g.,  $\beta = 0.5$ ); the stan-  
<sup>2154</sup> dard deviation of this prior would encode how certain we are about our  
<sup>2155</sup> prior beliefs. And, if we have no prior beliefs about the direction of the  
<sup>2156</sup> treatment effect, but we think it is unlikely to be very large, we could  
<sup>2157</sup> specify a normal prior centered at 0, which has the effect of "shrinking"  
<sup>2158</sup> the estimates closer to 0.<sup>11</sup>

<sup>2159</sup> As in our example above, maximum likelihood estimates and Bayesian

<sup>10</sup> This equation says that the probability of any value of  $\beta$  is "proportional to" 1, meaning that it's constant ("flat") regardless of what value  $\beta$  takes.

<sup>11</sup> The measures of variability that we discuss here account for statistical uncertainty reflecting the fact that we have only a finite sample size. If the sample size were infinite, there would be no uncertainty of this kind. Statistical uncertainty is only one kind of uncertainty, though. A more holistic view of the overall credibility of an estimate should also account for other things outside of the model, like study design issues and bias.

2160 estimates are going to be pretty similar if we have a lot of data or weak  
2161 priors. They will only diverge when we have strong priors or relatively  
2162 little data. The reason we are setting up these two different frameworks,  
2163 however, is that they provide very different inferential tools, as we'll see  
2164 in the next chapter.

2165 *5.2.2 Measures of effect size*

2166 Once we have measured something, we need to make a decision about  
2167 how to describe this effect to others. Sometimes we are working with  
2168 fairly intuitive relationships that are easy to describe. A researcher might  
2169 say, for example, that people who received milk-first tea drank the tea,  
2170 on average, five minutes quicker than people who received tea-first tea  
2171 (i.e., that  $\hat{\beta} = 5$  minutes). Time is measured in units like minutes and  
2172 seconds and so we all have a shared understanding of what five minutes  
2173 means.

2174 But what about our participants' ratings of tea quality, which were pro-  
2175 vided on an arbitrary 7-point rating scale that we devised? What does it  
2176 mean to that participants who drank milk-first tea rated it 1 point higher  
2177 than participants who drank tea-first tea (i.e., that  $\hat{\beta} = 1$  point)? And  
2178 how is this difference comparable to, for instance, a 1-point change on

<sup>2179</sup> a scale that has similar anchors (“terrible” and “delicious”) but uses a  
<sup>2180</sup> 100-point rating system?

<sup>2181</sup> To provide a common language for describing these relationships, some  
<sup>2182</sup> researchers use **standardized effect sizes**. A common standardized effect  
<sup>2183</sup> size is Cohen’s  $d$ , which provides a standardized estimate of the differ-  
<sup>2184</sup> ence between two means. There are many different ways to calculate  
<sup>2185</sup> Cohen’s  $d$  (Lakens 2013), but all approaches are usually some variant of  
<sup>2186</sup> the following formula:

$$d = \frac{\theta_M - \theta_T}{\sigma_{\text{pooled}}}$$

<sup>2187</sup> where the difference between means ( $\theta_T$  and  $\theta_M$ ) is divided by the  
<sup>2188</sup> pooled standard deviation  $\sigma_{\text{pooled}}$ . Intuitively, what you’re doing is  
<sup>2189</sup> taking the study effect ( $\beta$ ) and dividing it—scaling it—by the variation  
<sup>2190</sup> we saw between individuals in the study.

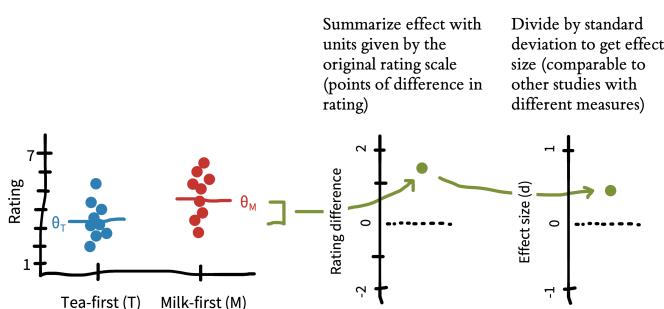


Figure 5.10  
Schematic effect size computation.

<sup>2191</sup> Let’s compute this measure for our tea-drinking study. We can just plug

<sup>2192</sup> in the estimates we see in figure 5.9 and compute the standard deviation

<sup>2193</sup> of our observed data:

$$\hat{d} = \frac{\hat{\theta}_M - \hat{\theta}_T}{\hat{\sigma}_{\text{pooled}}} = \frac{4.5 - 3.5}{1.25} = \frac{1}{1.25} = 0.80$$

<sup>2194</sup> In other words, the effect size of the difference between the two con-

<sup>2195</sup> ditions is 0.8 standard deviations. This process is shown graphically in

<sup>2196</sup> figure 5.10.<sup>12</sup>

<sup>2197</sup> We previously said that people who drank milk-first tea had quality

<sup>2198</sup> ratings that were, on average, 1 point higher on a 7-point scale ( $\beta = 1$

<sup>2199</sup> point). Cohen's  $d$  translates the arbitrary units of our rating scale into

<sup>2200</sup> a unit-less effect size that is measured in terms of the variation in the

<sup>2201</sup> data. You may find yourself wondering: "Why would I ever describe

<sup>2202</sup> things in terms of standard deviations?" The key benefit is that it allows

<sup>2203</sup> us to compare the size of the effect between studies that use different

<sup>2204</sup> measures.

<sup>2205</sup> Let's say that we ran a replication of our tea study with two changes: (1)

<sup>2206</sup> we studied patrons in a US cafe instead of a UK cafe, and (2) we used

<sup>2207</sup> a 100-point quality rating scale instead of a 7-point scale. Imagine that,

<sup>2208</sup> just as we found that participants in the UK rated the milk-first tea 1

<sup>2209</sup> point higher on a 7-point quality scale, US participants rated the milk-

<sup>12</sup> Cohen's  $d$ , also referred to as a **standardized mean difference (SMD)**, can be tricky to apply to more complex experimental designs, such as when you have within-participant designs and multiple measurements of each participant. For some guidance on this topic, see Lakens (2013).

2210 first tea 1 point higher on a *100-point* quality scale. It seems clear that  
2211 these effects are different because of the difference in scale. But how  
2212 different?

2213 It might at first seem reasonable just to normalize by the length of the  
2214 scale. So, maybe the UK experimental participants showed a 1/7 rating  
2215 effect and the US participants showed a 1/100 rating effect. The trouble  
2216 with this move is that it presupposes that participants from two differ-  
2217 ent populations are using two different scales in exactly the same way!  
2218 For example, maybe US participants made very clumpy judgments that  
2219 were mostly centered around 50 (perhaps because of a lack of milk tea  
2220 experience). Standardized effect sizes get around this kind of issue by  
2221 scaling according to the variability of the data.

2222 Let's compute the effect size for the cross-cultural replication. We'll  
2223 imagine that participants who drank milk-first tea gave an average rat-  
2224 ing of 50/100 and participants who drank tea-first tea rated it 49 on  
2225 average. But if their variability was also relatively lower, perhaps the  
2226 standard deviation of their ratings was only 5. Using the formula above,  
2227 we find

$$\hat{d}_{US} = \frac{\hat{\theta}_M - \hat{\theta}_T}{\hat{\sigma}_{\text{pooled}}} = \frac{50 - 49}{5} = \frac{1}{5} = 0.2$$

2228 A Cohen's  $d$  of 0.2 means that US cafe patrons rated their tea 0.2 stan-  
2229 dard deviations higher when it was milk-first, much smaller than the  
2230 0.8 standard deviation difference in the UK patrons.

2231 There are no hard and fast rules for interpreting what makes a big ef-  
2232 fect or a small effect, but people often refer back to a standard suggested  
2233 by Cohen (1992). On those standards,  $d = 0.8$  is a "large effect" and  
2234  $d = 0.2$  is a "small effect." But these effect size interpretation norms are  
2235 somewhat arbitrary. The key point here was that US and UK patrons  
2236 had the same raw score change in quality ratings ( $\hat{\beta} = 1$ ), and standard-  
2237 izing the differences allowed us to communicate that the difference was  
2238 larger among the UK patrons.

2239 Cohen's  $d$  is one of many standardized effect sizes that researchers can  
2240 use. Just as Cohen's  $d$  standardizes differences in group means, there  
2241 are also generalizations that allow for continuous treatment variables or  
2242 covariate adjustment (e.g., Pearson's  $r$ , as we discuss below;  $r^2$ ; or  $\eta^2$ ).  
2243 And there is a whole other set of effect-size measures for relationships  
2244 between binary variables (e.g., odds ratio). We'll be using effect sizes  
2245 throughout the book, but we'll be using Cohen's  $d$  as our example.<sup>13</sup>

<sup>13</sup> If you'd like to learn more about other varieties of effect size, take a look at Fritz, Morris, and Richler (2012) and Lakens (2013).

2246 *5.2.3 Pros and cons of standardizing effect sizes*

2247 Standardizing effect size helps communicate that a 1-point change on  
2248 a 7-point scale is not the same as a 1-point change on a 100-point scale.  
2249 But is it any better to say that the first change represents a 0.8 standard  
2250 deviation difference and the second a 0.08 standard deviation differ-  
2251 ence?

2252 Effect sizes allow us to compare results across studies more easily. Across  
2253 studies, researchers use different measures, different study designs, and  
2254 different populations. Standardization gives us a “common language”  
2255 to describe estimated relationships in these varied contexts. This lan-  
2256 guage is helpful when we want to aggregate and compare effects across  
2257 studies via meta-analysis. And it is also helpful when planning new stud-  
2258 ies. When trying to figure out how many participants to run in a study,  
2259 almost all techniques for sample size planning use standardized effect  
2260 sizes to determine how much data would be needed to reliably detect  
2261 an effect.

2262 Standardizing effect sizes has limitations, though. For example, if two  
2263 interventions produce the same absolute change in the same outcome  
2264 measure, but are studied in different populations in which the variabil-  
2265 ity on the outcome differs substantially, the interventions would pro-

<sup>2266</sup> duce different standardized mean differences (Baguley 2009) (see the

<sup>2267</sup> DEPTH box “Reliability paradoxes!” in chapter 8).

<sup>2268</sup> Imagine we conducted our tea experiment again, but this time with

<sup>2269</sup> decaf tea and focusing on children. Maybe milk-first tea tastes the same

<sup>2270</sup> amount better than tea-first tea for kids and for adults. But kids are, as a

<sup>2271</sup> rule, more variable in their responding than adults. This higher level of

<sup>2272</sup> variability would lead us to observe a smaller effect size in kids vs adults.

<sup>2273</sup> Recall that our UK adult SD was 1.25 and our effect size was  $d = 0.8$ .

<sup>2274</sup> Imagine that children’s SD is 2.5. In this scenario, even if tea led to the

<sup>2275</sup> same 1-point absolute change in ratings among adults and children, the

<sup>2276</sup> standardized effect size for kids would look half as big:

$$\hat{d}_{kids} = \frac{\hat{\theta}_M - \hat{\theta}_T}{\hat{\sigma}_{pooled}} = \frac{5 - 4}{2.5} = \frac{1}{2.5} = 0.4$$

<sup>2277</sup> This example highlights some of the challenges with standardization. If

<sup>2278</sup> we focused on the fact that both adults and children show a 1-point

<sup>2279</sup> change in ratings levels ( $\hat{\beta} = 1$ ), we would conclude that milk-first

<sup>2280</sup> tea ordering is as much better for adults as kids. If we focused on the

<sup>2281</sup> standardized effect sizes, however, we would conclude that the milk

<sup>2282</sup> ordering effect is twice as big for adults.

<sup>2283</sup> So which is better: Describing raw measures or standardized effect sizes?

2284 In general, our response is “Why not both?” But if you wanted to pick  
2285 one or the other, we recommend considering what type of measure-  
2286 ment you are using. With measures that yield common measurement  
2287 units that are likely to be reported in many studies already, use raw scores  
2288 (Baguley 2009). For example, if your study uses physical units such as  
2289 milliseconds (e.g., for reaction times) or counts (e.g., for a study track-  
2290 ing an outcome like number of words), these measurements can be quite  
2291 useful to compare across studies. Reporting raw measurements also can  
2292 allow you to check whether your measurements make sense—for exam-  
2293 ple, a reaction time of 70 milliseconds is inhumanly fast while a reaction  
2294 time of 10 seconds might be extremely slow (at least, for many speeded  
2295 tasks).

2296 In contrast, we recommend using standardized effect sizes for cases  
2297 where the measurement is relatively unlikely to be comparable with  
2298 other studies in its original form, or unlikely to be meaningful on its  
2299 own. For example, reporting the effect of an intervention on raw math  
2300 test scores is only meaningful if the reader knows how many items are  
2301 on the test, how difficult it is, and so forth. In such a case where it is  
2302 hard for a reader to be “calibrated” to the specific measurement units  
2303 you are using, standardized effect sizes may be the best way to report  
2304 your finding (Kelley and Preacher 2012).

2305 *5.3 Estimating the relationship between variables*

2306 Our focus up until now has been in estimating individual effects, but  
2307 sometimes we also want to estimate the relationship between two dif-  
2308 ferent variables. Extending our example, figure 5.11 shows the relation-  
2309 ship between the age of the tea taster and their rating of milk-first tea.  
2310 It seems that younger people overall like tea less than older people.<sup>14</sup>  
2311 How could we quantify this result?

2312 The first concept we need is **covariance**. Covariance captures the de-  
2313 gree to which we expect two variables to deviate from their means in  
2314 the same direction. We're looking at milk-first tea ratings  $M$  and partic-  
2315 ipant ages  $A$ . We can write the covariance between these two variables  
2316 as

$$\text{Cov}(M, A) = E[(M - \theta_M)(A - \theta_A)]$$

2317 This formula expresses the expected product of how much each ob-  
2318 servation differs from its expectation (mean) along each variable. Fig-  
2319 ure 5.12 shows these differences, which are multiplied together for each  
2320 point to get the covariation.<sup>15</sup>

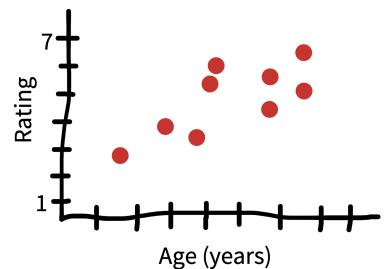


Figure 5.11  
The relationship between age and milk-first tea rating.

<sup>14</sup> Remember, this is a correlational relationship, and there's no causal inference possible here.

<sup>15</sup> This looks a little tricky, but it's actually very related to the basic concepts we've already seen. Remember when we introduced the standard deviation, we described it as the expected distance between new samples from a distribution and the mean of that distribution. The covariance is very related: the standard deviation is just  $\sqrt{\text{Cov}(X, X)}$ , in other words, the square root of the covariance of a variable with itself.

2321 This covariance number gives us an estimate of how much age and rat-  
 2322 ings covary, but its units are a bit funny: it's hard to know what to make  
 2323 of an expected deviation of 1 point-year. We can do a simple trick to  
 2324 standardize its units and make it into a wonderful form of effect size  
 2325 called the **correlation coefficient** (denoted  $r$ ). Remember that to create  
 2326 effect sizes above, we divided by the standard deviation of the variable.  
 2327 Here all we have to do is divide by the standard deviation of both vari-  
 2328 ables.

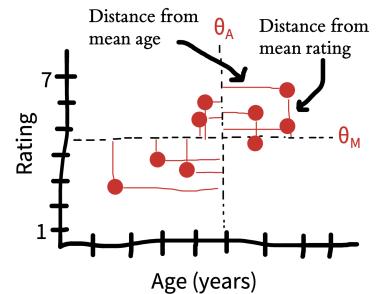


Figure 5.12  
 Estimating the covariation between age and milk-first tea rating.

$$r_{M,A} = \frac{Cov(M, A)}{\sigma_M \sigma_A}$$

2329 In other words, the correlation between two variables is the standard-  
 2330 ized covariation.

2331 The correlation coefficient is the most ubiquitous measure of associa-  
 2332 tion between variables. It ranges between -1, where two variables co-  
 2333 vary in exactly the opposite direction, to 1, when two variables covary  
 2334 perfectly. A correlation means that there is no association between two  
 2335 variables. A correlation of -1 or 1 doesn't mean that these two vari-  
 2336 ables have the same scale, however: it just means that they "move to-  
 2337 gether."

2338 Critically, a correlation is an effect size. Correlations can be compared

This section has described one way of looking at a correlation coefficient: as a standardized covariation. For a great discussion of all the different ways of

2339 across different measures and different studies (including both experi-  
2340 mental and observational studies), making it a very valuable scale-free  
2341 comparison tool.

2342 *5.4 Chapter summary: Estimation*

2343 In this chapter, we introduced the idea of estimating both individual  
2344 measurements and treatment effects from observed data. These ideas are  
2345 simple but they lay the foundations for hypothesis testing and modeling  
2346 (our next two chapters). Further, we set up the distinction between  
2347 Bayesian and frequentist approaches, which we will expand in the next  
2348 chapter since these traditions provide different inferential tools.



#### DISCUSSION QUESTIONS

1. In this chapter you learned about estimation, and in this book more generally, we have argued that the goal of an experiment is to provide a maximally precise estimate of a causal effect. Psychology as a field has often been criticized for focusing too much on inference and too little on estimation. Find an article in the journal *Psychological Science* that reports on an experiment or series of experiments and read the abstract. Does it mention an estimate of any particular quantity? What might be the benefits of reporting estimates in the study abstract?

2. Try the same exercise with a paper in the *New England Journal of*

*Medicine* or *Journal of the American Medical Association*. Find a paper and check if there is a mention of any specific quantity being estimated. (We suspect there will be!) Consider this contrast between the medical article and the psychology article. What do you make of this difference between fields?

2350

## READINGS

- A great narrative introduction to the history and practice of statistics:  
Salsburg, Daniel. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Macmillan.
- An open-source statistics textbook that follows a similar approach as chapters 5–7: Poldrack, Russell A. (2024). *Statistical Thinking for the 21st Century*. Available free online at <https://statsthinking21.org>.

2351

## References

- Baguley, Thom. 2009. “Standardized or Simple Effect Size: What Should Be Reported?” *British Journal of Psychology* 100 (3): 603–617.
- Cohen, Jacob. 1992. “A Power Primer.” *Psychological Bulletin* 112 (1): 155–159.
- Dunn, Peter M. 1997. “James Lind (1716–94) of Edinburgh and the Treatment of Scurvy.” *Archives of Disease in Childhood-Fetal and Neonatal Edition* 76 (1): F64–F65.
- Fritz, Catherine O., Peter E. Morris, and Jennifer J. Richler. 2012. “Effect

2353

Size Estimates: Current Use, Calculations, and Interpretation.” *Journal of Experimental Psychology: General* 141 (1): 2.

Kelley, Ken, and Kristopher J. Preacher. 2012. “On Effect Size.” *Psychological Methods* 17 (2): 137.

Kennedy, Maeve. 2003. “How to Make a Perfect Cuppa: Put Milk in First.” *How to Make a Perfect Cuppa: Put Milk in First.* <https://www.theguardian.com/uk/2003/jun/25/science.highereducation>.

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.

Lee Rodgers, Joseph, and W. Alan Nicewander. 1988. “Thirteen Ways to Look at the Correlation Coefficient.” *The American Statistician* 42 (1): 59–66.

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86 (3): 532–565.

Peirce, Charles Sanders, and Joseph Jastrow. 1884. “On Small Differences in Sensation.” *Memoirs of the National Academy of Sciences* 3.

Poldrack, Russell A. 2024. *Statistical Thinking for the 21st Century.* <https://statsthinking21.org/>.

Salsburg, David. 2002. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* Macmillan.

## 2355 6 INFERENCE



### LEARNING GOALS

- Discuss the purpose of statistical inference
- Define  $p$ -values and Bayes Factors
- Consider common fallacies about inference (especially for  $p$ -values)
- Reason about sampling variability
- Define and reason about confidence intervals

2356

2357 We've been arguing that experiments are about measuring effects. The  
2358 effects we are interested in are causal effects for a group of people, but  
2359 that group is almost always bigger than the participants in an experi-  
2360 ment. **Statistical inference** is the process of going beyond the specific  
2361 characteristics of the sample that you measured to make generalizations  
2362 about the broader population.

2363 Chapter 5 already showed us how to make one simple inference: esti-  
2364 mating population parameters using both frequentist and Bayesian tech-

2365 niques. Estimating population parameters is an important first step. But  
2366 often we want to make more sophisticated inferences so that we can an-  
2367 swer questions such as:

- 2368 1. How likely is it that this pattern of measurements was produced  
2369 by chance variation?
- 2370 2. Do these data provide more support for one hypothesis or an-  
2371 other?
- 2372 3. How precise is our estimate of an effect?
- 2373 4. What portion of the variation in the data is due to a particular ma-  
2374 nipulation (as opposed to variation between participants, stimulus  
2375 items, or other manipulations)?

2376 Question (1) is associated with one particular type of statistical infer-  
2377 ence method—**null hypothesis significance testing** (NHST) in the fre-  
2378 quentist statistical tradition. NHST has become synonymous with data  
2379 analysis, such that in the vast majority of research papers (and research  
2380 methods courses), all of the reported analyses are tests of this type. Yet,  
2381 this equivalence is quite problematic.

2382 The move to “go test for significance” before visualizing your data and  
2383 trying to understand sources of variation (participants, items, manipula-  
2384 tions, and other sources) is one of the most unhelpful strategies for an

2385 experimenter. Whether  $p < 0.05$  or not, a test of this sort gives you lit-  
 2386 erally *one bit* of information about your data.<sup>1</sup> Considering effect sizes  
 2387 and their variation more holistically, including using the kinds of visu-  
 2388 alizations we advocate in chapter 15, gives you a much richer sense of  
 2389 what happened in your experiment!

2390 In this chapter, we will describe NHST, the conventional method that  
 2391 many students still learn (and many scientists still use) as their primary  
 2392 method for engaging with data. All practicing experimentalists need  
 2393 to understand NHST, both to read the literature and also to apply this  
 2394 method in appropriate situations. For example, NHST may be a rea-  
 2395 sonable tool for testing whether an intervention leads to a difference  
 2396 between a treatment condition and an appropriate control. But we will  
 2397 also try to contextualize NHST as a very special case of a broader set of  
 2398 statistical inference strategies. Further, we will continue to flesh out our  
 2399 account of how some of the pathologies of NHST have been a driver  
 2400 of the replication crisis.

2401 If NHST approaches have so many issues, what should replace them?

2402 Figure 6.1 shows one way of organizing different inferential approaches.

2403 There has been a recent move toward the use of Bayes Factors to quan-  
 2404 tify the evidence in support of different candidate hypotheses. Bayes  
 2405 Factors can help answer questions like (2). We introduce these tools and

<sup>1</sup> In the information theoretic sense, as well as the common sense!

	Frequentist	Bayesian
Measurement focused	estimate with confidence interval	posterior distribution with credible interval
Hypothesis focused	$p$ value from null hypothesis significance test	Bayes factor

Figure 6.1

Clarifying the distinctions between Bayesian and Frequentist paradigms and the tools they offer for measurement and hypothesis testing. For many settings, we think the measurement mindset is more useful. Adapted from Kruschke and Liddell (2018b).

2406 we believe that they have broader applicability than the NHST frame-  
2407 work and should be known by students. On the other hand, Bayes Fac-  
2408 tors are not a panacea. They have many of the same problems as NHST  
2409 when they are applied dichotomously.

2410 Instead of dichotomous frequentist or Bayesian hypothesis testing, we  
2411 follow our thematic emphasis on MEASUREMENT PRECISION and advo-  
2412 cate for a **measurement** strategy, which is more suited toward questions  
2413 (3) and (4) (Cumming 2014; Kruschke and Liddell 2018b). The goal of  
2414 these strategies is to yield an accurate and precise estimate of the rela-  
2415 tionships underlying observed variation in the data.

2416 This isn't a statistics book, and we won't attempt to teach the full array  
2417 of important statistical concepts that will allow students to build good  
2418 models of a broad array of datasets (Sorry!).<sup>2</sup> But we do want you to  
2419 be able to reason about inference and modeling. In this chapter, we'll  
2420 start by making some inferences about our tea-tasting example from the  
2421 last chapter, using this example to build up intuitions about hypothesis  
2422 testing and inference. Then, in chapter 7, we'll start to look at more  
2423 sophisticated models and how they can be fit to real datasets.

<sup>2</sup> If you're interested in going deeper, here are two books that have been really influential for us. The first is Gelman and Hill (2006) and its successor Gelman, Hill, and Vehtari (2020), which teach regression and multi-level modeling from the perspective of data description. The second is McElreath (2018), a course on building Bayesian models of the causal structure of your data. Honestly, neither is an easy book to sit down and read (unless you are the kind of person who reads statistics books on the subway for fun), but both really reward detailed study. We encourage you to get together a reading group and go through the exercises in one of these together. It'll be well worthwhile in its impact on your statistical and scientific thinking.

## 2424 6.1 Sampling variation

2425 In chapter 5, we introduced Fisher's tea-tasting experiment and dis-  
 2426 cussed how to estimate means and differences in means from our ob-  
 2427 served data. These so-called point estimates represent our best guesses  
 2428 about the population parameters given the data—and possibly also given  
 2429 our prior beliefs. We can also report how much statistical uncertainty  
 2430 is involved in these point estimates.<sup>3</sup> Quantifying and reasoning about  
 2431 this uncertainty is an important goal: in our original study we only had  
 2432 nine participants in each group, which will only provide a low-precision  
 2433 (i.e., highly uncertain) estimate of the population. By contrast, if we re-  
 2434 peated the experiment with 200 participants in each group, the data  
 2435 would be far less noisy, and we would have much less uncertainty, even  
 2436 if the point estimates happened to be identical.

### 2437 6.1.1 Standard errors

2438 To characterize the uncertainty in an estimate, it helps to picture its  
 2439 **sampling distribution**, which is the distribution of the estimate across  
 2440 different, hypothetical, samples. That is, let's imagine that we con-  
 2441 ducted the tea experiment not just once but dozens, hundreds, or even  
 2442 thousands of times. This idea is often called **repeated sampling** as a short-  
 2443 hand. For each hypothetical sample, we use similar recruitment meth-

<sup>3</sup> As in the previous chapter, we're only capturing *statistical* uncertainty. A holistic view of a particular estimate's credibility also include everything else you know about the study design.

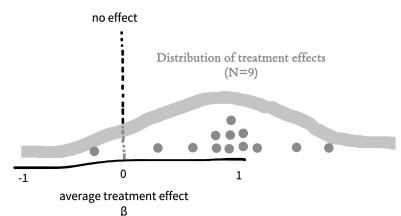


Figure 6.2  
 Sampling distribution for the treatment effect in the tea-tasting experiment, given many different repetitions of the same experiment, each with  $n = 9$  per group. Circles represent average treatment effects from different individual experiments, while the thick line represents the form of the underlying distribution.

ods to recruit a new sample of participants, and we compute  $\hat{\beta}$  for that sample. Would we get exactly the same answer each time? No, simply because the samples will have some random variability (noise). If we plotted these estimates,  $\hat{\beta}$ , we would get the sampling distribution in figure 6.2.

### CODE

In this chapter and the subsequent statistics and visualization chapters of the book, we'll try to facilitate understanding and illustrate how to use these concepts in practice by giving the R code we use in constructing our examples in these code boxes. We'll assume that you have some knowledge of base R and the Tidyverse—to get started with these, go ahead and take a look at appendix D if you haven't already. Although our figures are often drawn by hand, even the hand-drawn ones are based on actual simulation results!

Since we're going to be working with lots of data from the tea tasting example, we wrote a function called `make_tea_data()` that creates a `tibble` with some (made-up) data from our modern tea-tasting experiment. You can find the function on GitHub ([https://github.com/langcog/experimentology/blob/main/helper/tea\\_helper.qmd](https://github.com/langcog/experimentology/blob/main/helper/tea_helper.qmd)) if you want to follow along.

```
tea_data <- make_tea_data(n_total = 18)
```

2450 Now imagine we also did thousands of repetitions of the experiment  
 2451 with  $n = 200$  per group instead of  $n = 9$  per group. Figure 6.3 shows  
 2452 what the sampling distribution might look like in that case. Notice how  
 2453 much narrower the sampling distribution becomes when we increase  
 2454 the sample size, showing our decreased uncertainty. More formally, the  
 2455 standard deviation of the sampling distribution itself, called the **standard**  
 2456 **error**, decreases as the sample size increases.

2457 The sampling distribution is not the same thing as the distribution of tea  
 2458 ratings in a single sample. Instead, it's a distribution of *estimates across*  
 2459 *samples of a given size*. In essence, it tells us what the mean of a new  
 2460 experiment might be, if we ran it with a particular sample size.

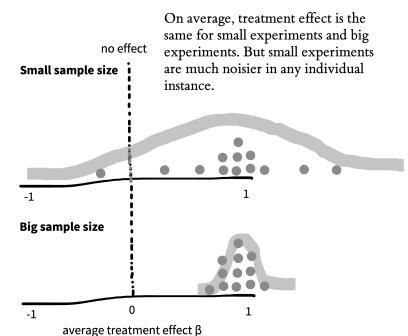


Figure 6.3  
 Comparing sampling distributions for the treatment effect with smaller and larger size samples.

## CODE

To do simulations where we repeat the tea-tasting experiment over and over again, we're using a special tidyverse function from the `purrr` library:

`map()`. `map()` is an extremely powerful function that allows us to run another function (in this case, the `make_tea_data()` function that we introduced last chapter) many times with different inputs. Here we create a tibble made up of a set of 1,000 runs of the `make_tea_data()` function.

```
samps <- tibble(sim = 1:1000) |>  
  
  mutate(data = map(sim, \((i) make_tea_data(n_total = 18)))) |>  
  unnest(cols = data)
```

Next, we just use the `group_by()` and `summarize()` workflow from appendix D to get the estimated treatment effect for each of these simulations.

```
tea_summary <- samps |>  
  group_by(sim, condition) |>  
  summarize(mean_rating = mean(rating)) |>  
  group_by(sim) |>  
  summarize(delta = mean_rating[condition == "milk first"] -  
            mean_rating[condition == "tea first"])
```

This tibble gives us what we would need to plot the sampling distributions above in figure 6.2 and figure 6.3.

2462

### 2463 6.1.2 *The central limit theorem*

2464 We talked in the last chapter about the normal distribution, a conve-  
2465 nient and ubiquitous tool for quantifying the distribution of measure-  
2466 ments. A shocking thing about sampling distributions for many kinds  
2467 of estimates—and for *all* maximum likelihood estimates—is that they  
2468 become normally distributed as the sample size gets larger and larger.  
2469 This result holds even for estimates that are not even remotely normally  
2470 distributed in small samples!

2471 For example, say we are flipping a coin and we want to estimate the  
 2472 probability that it lands heads ( $p_H$ ). If we draw samples each consisting  
 2473 of only  $n = 2$  coin flips, figure 6.4 is the sampling distribution of the  
 2474 estimates ( $\hat{p}_H$ ). This sampling distribution doesn't look normally dis-  
 2475 tributed at all—it doesn't have the characteristic "bell curve" shape! In  
 2476 a sample of only two coin flips,  $\hat{p}_H$  can only take on the values 0, 0.5, or  
 2477 1.

2478 But look what happens as we draw increasingly larger samples in fig-  
 2479 ure 6.5: we get a normal distribution! This tendency of sampling distri-  
 2480 butions to become normal as  $n$  becomes very large reflects a deep and  
 2481 elegant mathematical law called the **central limit theorem**.

2482 The practical upshot is that the central limit theorem directly helps us  
 2483 characterize the uncertainty of sample estimates. For example, when  
 2484 the sample size is reasonably large (approximately  $n > 30$  in the case  
 2485 of sample means) the standard error (i.e., the standard deviation of the  
 2486 sampling distribution) of a sample mean is approximately  $\widehat{SE} = \sigma / \sqrt{n}$ .  
 2487 The sampling distribution becomes narrower as the sample size in-  
 2488 creases because we are dividing by the square root of the number of  
 2489 observations.

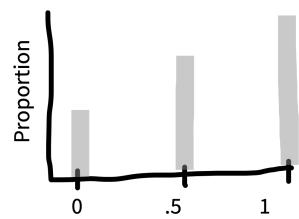


Figure 6.4  
Sampling distribution of samples from a biased coin ( $n = 2$  flips per sample). Bar height is the proportion of flips resulting in a particular mean.

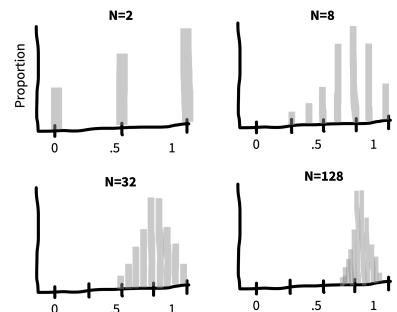


Figure 6.5  
Sampling distribution for 2, 8, 32, and 128 flips.

 CODE

Even though our figures are hand-drawn, they're based on real simulations. For our central limit theorem simulations, we again use the `map()` function. We set up a tibble with the different values we want to try (which we call `n_flips`). Then we make use of the `map()` function to run `rbinom()` (random binomial samples) for each value of `n_flips`.

One trick we make use of here is that `rbinom()` takes an extra argument that says how many of these random values you want to generate. Here we generate `nsamps = 1000` samples, giving us 1,000 independent replicates at each `n`. But returning an array of 1,000 values for a single value of `n_flips` results in something odd: the value for each element of `flips` is an array. To deal with that, we use the `unnest()` function, which expands the array back into a normal tibble.

```
n_samps <- 1000

n_flips_list <- c(2, 8, 32, 128)

sample_p <- tibble(n_flips = n_flips_list) |>
  mutate(flips = map(n_flips, \f) rbinom(n = n_samps, size = f, prob = .7))) |>
  unnest(cols = flips) |>
  mutate(p = flips / n_flips)
```

## 2491 6.2 From variation to inference

2492 Let's go back to Fisher's tea-tasting experiment. The first innovation  
2493 of that experiment was the use of randomization to recover an estimate  
2494 of the causal effect of milk ordering. But there was more to Fisher's  
2495 analysis than we described.

2496 The second innovation of the tea-tasting experiment was the idea of  
2497 creating a model of what might happen during the experiment. Specif-  
2498 ically, Fisher described a hypothetical **null model** that would arise if the  
2499 lady had chosen cups by chance rather than because of some tea sen-  
2500 sitivity. In our tea-rating experiment, the null model describes what  
2501 happens when there is no difference in ratings between tea-first and  
2502 milk-first cups. Under the null model, the true treatment effect ( $\beta$ ) is  
2503 zero.

2504 Even with an actual treatment effect of zero, across repeated sampling,  
2505 we should see some variation in  $\hat{\beta}$ , our *estimate* of the treatment effect.  
2506 Sometimes we'll get a small positive effect, sometimes a small negative  
2507 one. Occasionally just by chance we'll get a big effect. This is just sam-  
2508 pling variation as we described above.

2509 Fisher's innovation was to quantify the probability of observing vari-  
2510 ous values of  $\hat{\beta}$ , given the null model. Then, if the observed data that

2511 were very low probability under the null model, we could declare that  
2512 the null was rejected. How unlikely must the observed data be in or-  
2513 der to reject the null? Fisher declared that it is “usual and convenient  
2514 for experimenters to take 5 percent as a standard level of convenience,”  
2515 establishing the 0.05 cutoff that has become gospel throughout the sci-  
2516 ences.<sup>4</sup>

2517 Let’s take a look at what the null model might look like. We already  
2518 tried out repeating our tea-tasting experiment thousands of times in our  
2519 discussion of sampling above. Now in figure 6.6, we do the same thing  
2520 but we assume that the **null hypothesis** of no treatment effect is true.  
2521 The plot shows the distribution of treatment effects  $\hat{\beta}$  we observe: some  
2522 a little negative, some a little positive, and a few substantially positive  
2523 or negative, but mostly zero.

2524 Let’s apply the  $p < 0.05$  standard. If our observation has less than a 5%  
2525 probability under the null model, then the null model is likely wrong.  
2526 The red dashed lines on figure 6.6 show the point below which only  
2527 2.5% of the data are found and the point above which only 2.5% of the  
2528 data are found. These are called the **tails** of the distribution. Because  
2529 we’d be equally willing to accept milk-first tea or tea-first tea being bet-  
2530 ter, we consider both positive and negative observations as possible.<sup>5</sup>

<sup>4</sup> Actually, right after establishing 0.05 as a cutoff, Fisher then writes that “in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon … in order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.” In other words, Fisher was all for replication!

<sup>5</sup> Because we’re looking at both tails of the distribution, this is called a “two-tailed” test.

## CODE

To simulate our null model, we can do the same kind of thing we did before, just specifying to our `make_tea_data()` function that the true difference in effects is zero!

```
n_sims <- 1000

null_model <- tibble(sim = 1:n_sims, n = 18) |>

  mutate(data = map(sim, \(i) make_tea_data(n_total = n, delta = 0))) |>

  unnest(cols = data)
```

Again we use `group_by()` and `summarize()` to get the distribution of treatment effects under the null hypothesis.

```
null_model_summary <- null_model |>

  group_by(sim, condition) |>

  summarize(mean_rating = mean(rating)) |>

  group_by(sim) |>

  summarize(delta = mean_rating[condition == "milk first"] - 

    mean_rating[condition == "tea first"])
```

2531

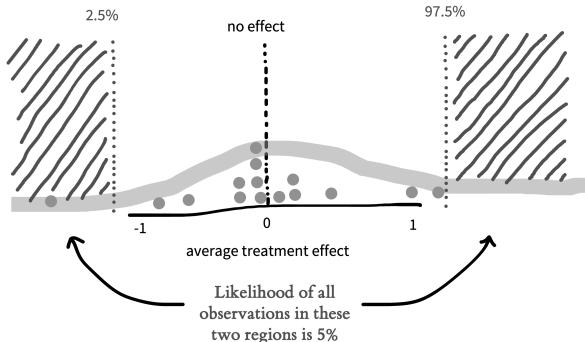


Figure 6.6

One example of the distribution of treatment effects under the null model (with  $n = 9$  per group). The red regions indicate the part of the distribution in which less than 5% of observations should fall.

2532 Figure 6.6 captures the logic of NHST: if the observed data fall in the  
2533 region that has a probability of less than 0.05 under the null model, then  
2534 we reject the null. So, then when we observe some particular treatment  
2535 effect  $\hat{\beta}$  in a single (real) instance of our experiment, we can compute  
2536 the probability of these data or any data more extreme than ours under  
2537 the null model.<sup>6</sup> This probability is our  $p$ -value, and if it is small, it  
2538 gives us license to conclude that the null is false.

2539 As we saw before, the larger the sample size, the smaller the standard  
2540 error. That's true for the null model too! figure 6.7 shows the expected  
2541 null distribution for a bigger experiment.

<sup>6</sup> The “more extreme” part deserves a little explanation. Any individual outcome is relatively unlikely by itself, just because it’s surprising that the estimate is that exact value (we’re simplifying here, it gets a bit trickier when you are talking about real numbers). What we care about instead is a *group* of values. The ones that are in the middle of the distribution are, considered as a group, quite likely; the ones on the tails are, as a group, less likely. We want to know if the probability of the group of datapoints that includes our observation and anything even further out on the tails is collectively less than 0.05.

2542 The more participants in the experiment, the tighter the null distribu-  
2543 tion becomes, and hence the smaller the region in which we should  
2544 expect a null treatment effect to fall. Because our expectation based on  
2545 the null becomes more precise, we will be able to reject the null based  
2546 on smaller treatment effects. In this type of hypothesis testing, as with  
2547 estimation, our goals matter. If we're merely testing a hypothesis out of  
2548 curiosity, perhaps we don't want to measure too many cups of tea. But  
2549 if we were designing the tea strategy for a major cafe chain, the stakes  
2550 would be higher; in that case, maybe we'd want to do a more extensive  
2551 experiment!

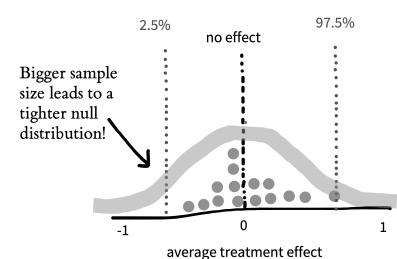


Figure 6.7  
Example distribution of treatment effects under the null model for a larger experiment.

### CODE

We can do a more systematic simulation of the null regions for different sample sizes by simply adding a parameter to our simulation.

```
n_sims <- 10000

null_model_multi_n <- expand_grid(sim = 1:n_sims, n = c(12, 24, 48, 96)) |>
  mutate(sim_data = map(n, \(n_i) make_tea_data(n_total = n_i, delta = 0))) |>
  unnest(cols = sim_data)

null_model_summary_multi_n <- null_model_multi_n |>
  group_by(n, sim, condition) |>
  summarize(mean_rating = mean(rating)) |>
  group_by(n, sim) |>
  summarize(delta = mean_rating[condition == "milk first"] -
            mean_rating[condition == "tea first"])

null_model_quantiles_multi_n <- null_model_summary_multi_n |>
  group_by(n) |>
  summarize(q_025 = quantile(delta, .025),
            q_975 = quantile(delta, .975))
```

Here is the plotting code to produce a comparable figure to our illustration:

```
ggplot(null_model_summary_multi_n, aes(x = delta)) +  
  facet_wrap(vars(n), nrow = 1, labeller = label_both) +  
  geom_histogram(binwidth = .25) +  
  geom_vline(xintercept = 0, color = pal$grey, linetype = "dotted") +  
  geom_vline(data = null_model_quantiles_multi_n,  
             aes(xintercept = q_025), color = pal$red, linetype = "dotted") +  
  geom_vline(data = null_model_quantiles_multi_n,  
             aes(xintercept = q_975), color = pal$red, linetype = "dotted") +  
  xlim(-2.5, 2.5) +  
  labs(x = "Difference in rating", y = "Frequency")
```

2554

2555 One last note: You might notice an interesting parallel between the  
2556 NHST paradigm and Popper's falsificationist philosophy (introduced in  
2557 chapter 2). In both cases, you never get to *accept* the actual hypothesis  
2558 of interest. The only thing you can do is observe evidence that is incon-  
2559 sistent with the null hypothesis. The added limitation of NHST is that  
2560 the only hypothesis you can falsify is the null!?

1989

### 2561 6.3 Making inferences

2562 In the tea-tasting example we were just considering, we were trying  
2563 to make an inference from our sample to the broader population. In  
2564 particular, we were trying to test whether milk-first tea was rated as

2565 better than tea-first tea. Our inferential goal was a clear, binary answer:

2566 Is milk-first tea better?

2567 By defining a  $p$ -value, we got one procedure for giving this answer. If

2568  $p < 0.05$ , we reject the null. Then we can look at the direction of the

2569 difference and, if it's positive, declare that milk-first tea is "significantly"

2570 better. Let's compare this procedure to a different process that builds on

2571 the Bayesian estimation ideas we described in the previous chapter. We

2572 can then come back to examine NHST in light of that framework.

### 2573 6.3.1 Bayes Factors

2574 Bayes Factors are a method for quantifying the support for one hypoth-

2575 esis over another, based on an observed dataset. They don't tell you the

2576 probability that a particular hypothesis is right, but they let you com-

2577 pare two different ones.

2578 Informally, we've now discussed two different distinct hypotheses

2579 about the tea situation: our participants could have *no* tea discrim-

2580 ination ability—leading to chance performance. We call this  $H_0$ .

2581 Or they could have some nonzero ability—leading to greater than

2582 chance performance. We call this  $H_1$ . The Bayes Factor is simply the

2583 likelihood of the data (in the technical sense used above) under  $H_1$

2584 vs. under  $H_0$  (figure 6.8). The Bayes Factor is a ratio, so if it is greater

$$BF = \frac{p(\text{data}|H_1)}{p(\text{data}|H_0)}$$

↓    ↓

Likelihood of data under hypothesis of non-zero difference in ability    Likelihood of data under null hypothesis of zero difference

Figure 6.8  
The Bayes Factor (BF).

2585 than 1, the data are more likely under  $H_1$  than they are under  $H_0$ —and  
 2586 vice versa for values between 1 and 0. A BF of 3 means there is three  
 2587 times as much evidence for  $H_1$  than  $H_0$ , or equivalently 1/3 as much  
 2588 evidence for  $H_0$  as  $H_1$ .<sup>8</sup>

### CODE

Bayes Factors are delightfully easy to compute using the BayesFactor R package (Morey and Rouder 2023). All we do is feed in the two sets of ratings to the `ttestBF()` function!

```
library(BayesFactor)

tea_bf <- ttestBF(x = filter(tea_data, condition == "milk first")$rating,
                    y = filter(tea_data, condition == "tea first")$rating,
                    paired = FALSE)
```

2589 There are a couple of things to notice about the Bayes Factor. The first  
 2590 is that, like a  $p$ -value, it is inherently a continuous measure. You can  
 2591 artificially dichotomize decisions based on the Bayes Factor by declaring  
 2592 a cutoff (say,  $\text{BF} > 3$  or  $\text{BF} > 10$ ), but there is no intrinsic threshold at  
 2593 which you would say the evidence is “significant.” Some guidelines for  
 2594 interpretation (from S. N. Goodman 1999) are shown in table 6.1.<sup>9</sup> On  
 2595 the other hand, cutoffs like  $\text{BF} > 5$  or  $p < 0.05$  are not very informative.  
 2596 So, although we provide this table to guide interpretation, we caution

<sup>8</sup> Sometimes people refer to the BF in favor of  $H_1$  as the  $BF_{10}$  and the BF in favor of  $H_0$  as the  $BF_{01}$ . This notation is a bit confusing because the first of these looks like the number 10.

<sup>9</sup> Some like the guidelines provided by Jeffreys (1961), which include categories such as “barely worth mentioning” ( $1 > \text{BF} > 3$ ).

2598 that you should always report and interpret the actual Bayes Factor, not

2599 whether it is above or below some cutoff.

Table 6.1  
S. N. Goodman (1999) interpretation guidelines for Bayes Factors.

BF range	Interpretation
< 1	Negative (supports $H_0$ )
1–5	Weak
5–10	Moderate
10–20	Moderate to strong
20–100	Strong to very strong

2600 The second thing to notice about the Bayes Factor is that it doesn't de-

2601 pend on our prior probability of  $H_1$  vs.  $H_0$ . We might think of  $H_1$  as

2602 very implausible. But the BF is independent of that prior belief. So that

2603 means it's a measure of how much the evidence should shift our beliefs

2604 away from our prior. One nice way to think about this is that the Bayes

2605 Factor computes how much our beliefs—whatever they are—should be

2606 changed by the data (Morey and Rouder 2011).

2607 In practice, the thing that is both tricky and good about Bayes Factors

2608 is that you need to define an actual model of what  $H_0$  and  $H_1$  are. That

2609 process involves making some assumptions explicit. We won't go into

2610 how to make these models here—this is a big topic that is covered ex-  
2611 tensively in books on Bayesian data analysis.<sup>10</sup> The goal here is just to  
2612 give a general sense of what Bayes Factors are.

### 2613 6.3.2 *p*-values

2614 Now let's turn back to NHST and the *p*-value. We already have a work-  
2615 ing definition of what a *p*-value is from our discussion above: it's the  
2616 probability of the data (or any data that would be more extreme) under  
2617 the null hypothesis. How is this quantity related to either our Bayesian  
2618 estimate or the BF? Well, the first thing to notice is that the *p*-value is  
2619 very close (but not identical) to the likelihood itself.<sup>11</sup>

2620 Next we can use a simple statistical test, a *t*-test, to compute *p*-values for  
2621 our experiment. In case you haven't encountered one, a *t*-test is a pro-  
2622 cedure for computing a *p*-value by comparing the distribution of two  
2623 variables using the null hypothesis that there is no difference between  
2624 them.<sup>12</sup> The *t*-test uses the data to compute a **test statistic** whose dis-  
2625 tribution under the null hypothesis is known. Then the value of this  
2626 statistic can be converted to *p*-values for making an inference.

<sup>10</sup> Two good ones beyond the McElreath book mentioned above are Gelman et al. (1995), which is a bit more statistical, and Kruschke (2014), which is a bit more focused on psychological data analysis. An in-prep web-book by Nicenboim et al. (<https://vasishth.github.io/bayescogsci/book>) also looks great.

<sup>11</sup> The likelihood—for both Bayesians and frequentists—is the probability of the data, just like the *p*-value. But unlike the *p*-value, it doesn't include the probability of more extreme data as well.

<sup>12</sup> *t*-tests can also be used in cases where one sample is being compared to some baseline.

## CODE

The standard `t.test()` function is built into R via the default `stats` package. Here we simply make sure to specify the variety of test we want by using the flags `paired = FALSE` and `var.equal = TRUE` (denoting the assumption of equal variances).

```
tea_t <- t.test(x = filter(tea_data, condition == "milk first")$rating,
                  y = filter(tea_data, condition == "tea first")$rating,
                  paired = FALSE, var.equal = TRUE)
```

2627

2628 Imagine we conduct a tea-tasting experiment with  $n = 48$  and perform  
 2629 a *t*-test on our experimental results. In this case, we see that the differ-  
 2630 ence between the two groups is significant at  $p < 0.05$ :  $t(46) = 2.86$ ,  
 2631  $p = 0.006$ .

2632 The expression  $t(46) = 2.86$ ,  $p = 0.006$  is the standard way to report  
 2633 a *t*-test according to the American Psychological Association. The first  
 2634 part of this report gives the *t* value, qualified by the **degrees of free-**  
 2635 **dom** for the test in parentheses. We won't focus much on the idea of  
 2636 degrees of freedom here, but for now it's enough to know that this num-  
 2637 ber quantifies the amount of information given by the data, in this case  
 2638 48 datapoints minus the two means (one for each of the samples).

2639 Let's compare *p* values and Bayes Factors (computed using the default

Table 6.2  
 Comparison of *p*-value and BF for several different (randomly-generated) tea-tasting scenarios.

N	Effect size	p-value	BF
12	0.5	> .999	0.5
12	1.0	.076	1.4
12	1.5	.002	18.7
24	0.5	.858	0.4
24	1.0	.061	1.5
24	1.5	.009	5.6
48	0.5	.002	17.7
48	1.0	.033	2.0

2640 setup in the BayesFactor R package). In table 6.2, the rows represent  
2641 simulated experiments with varying total numbers of participants (N)  
2642 and varying average treatment effects. Both  $p$  and BF go up with more  
2643 participants and larger effects. In general, BFs tend to be a bit more  
2644 conservative than  $p$ -values, such that  $p < 0.05$  can sometimes translate  
2645 to a BF of less than 3 (Benjamin et al. 2018). For example, take a look  
2646 at the row with 48 participants and an effect size of 1: the  $p$  value is less  
2647 than 0.05, but the Bayes Factor is only 2.0.

2648 The critical thing about  $p$ -values, though, is not just that they are a kind  
2649 of data likelihoods. It is that they are used in a *specific inferential procedure*.

2650 The logic of NHST is that we make a binary decision about the presence  
2651 of an effect. If  $p < 0.05$ , the null hypothesis is rejected; otherwise not.

2652 As Fisher (1949, p. 19) wrote,

2653 It should be noted that the null hypothesis is never proved  
2654 or established, but is possibly disproved, in the course of  
2655 experimentation. Every experiment may be said to exist  
2656 only in order to give the facts a chance of disproving the  
2657 null hypothesis.

2658 The main problem with  $p$ -values from a scientific perspective is that  
2659 researchers are usually interested in not just rejecting the null hypothesis

2660 but also in the evidence for the alternative (the one we are interested in).

2661 The Bayes Factor is one approach to quantifying positive evidence for

2662 the alternative hypothesis in a Bayesian framework. This issue with the

2663 Fisher approach to  $p$ -values has been known for a long time, though,

2664 and so there is an alternative frequentist approach as well.

### 2665 6.3.3 The Neyman-Pearson approach

2666 One way to “patch” NHST is to introduce a decision-theoretic view,

2667 shown in figure 6.9.<sup>13</sup> On this view, called the Neyman-Pearson view,

2668 there is a real  $H_1$ , albeit one that is not specified. Then the true state of

2669 the world could be that  $H_0$  is true or  $H_1$  is true. The  $p < 0.05$  criterion

2670 is the threshold at which we are willing to reject the null, and so this

2671 constitutes our **false positive rate**  $\alpha$ . But we also need to define a **false**

2672 **negative rate**, which is conventionally called  $\beta$ .<sup>14</sup>

2673 Setting these rates is a decision problem: If you are too conservative in

2674 your criteria for the intervention having an effect, then you risk a **false**

2675 **negative**, where you incorrectly conclude that it doesn’t work. And if

2676 you’re too liberal in your assessment of the evidence, then you risk a

2677 **false positive**.<sup>15</sup> In practice, however, people usually leave  $\alpha$  at 0.05 and

2678 try to control the **false negative rate** by increasing their sample size.

		Inference	
		Reject null ( $H_0$ )	Fail to reject null ( $H_0$ )
Reality	Null ( $H_0$ ) is true	False positive $\alpha$	Correct rejection $1 - \alpha$
	Null ( $H_0$ ) is false	True positive $1 - \beta$	False negative $\beta$

Figure 6.9  
Standard decision matrix for the Neyman-Pearson approach to statistical inference.

<sup>13</sup> A little bit of useful history here is given in Cohen (1990), and we also recommend Gigerenzer (1989) for a broader perspective.

<sup>14</sup> Unfortunately,  $\beta$  is very commonly used for regression coefficients—and for that reason we’ve used it as our symbol for causal effects. We’ll be using these  $\beta$ s in the next chapter as well. Those  $\beta$ s are not to be confused with false negative rates. Sorry, this is just a place where statisticians have used the same Greek letter for two different things.

2679 As we saw in figure 6.6, the larger the sample, the better your chance  
2680 of rejecting the null for any given non-null effect. But these chances  
2681 will depend also on the effect size you are estimating. This formula-  
2682 tion gives rise to the idea of classical power analysis, which we cover in  
2683 chapter 10. Most folks who defend binary inference are interested in  
2684 using the Neyman-Pearson approach. In our view, this approach has  
2685 its place (it's especially useful for power analysis) but it still suffers from  
2686 the substantial issues that plague all binary inference techniques.

<sup>15</sup> To make really rational decisions, you could couple this chart to some kind of utility function that assessed the costs of different outcomes. For example, you might think it's worse to proceed with an intervention that doesn't work than to stay with business as usual. In that case, you'd assign a higher cost to a false positive and accordingly try to adopt a more conservative criterion. We won't cover this kind of decision analysis here, but Pratt et al. (1995) is a classic textbook on statistical decision theory if you're interested.

 DEPTH

### *Nonparametric resampling under the null*

Hypothesis testing requires knowing the null distribution. In the examples above, it was easy to use statistical theory to work out the null distribution using knowledge of the binomial or normal distribution. But sometimes we don't know what the null distribution would look like. What if the ratings data from our tea-tasting experiment were very skewed, such that there were many low ratings and a few very high ratings (as in figure 6.10)?

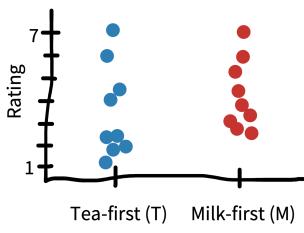


Figure 6.10

A small tea-tasting experiment with a skewed distribution of ratings.

With skewed data like these, we couldn't proceed with a  $t$ -test in good conscience because, with only  $n = 18$ , we can't necessarily trust that the central limit theorem has "kicked in" sufficiently for the test to work despite the skewness. Put another way, we can't be sure that the null distribution is normal (Gaussian) in this case.

An alternative way to approximate a null distribution is through nonparametric resampling. **Resampling** means that we're going to draw new samples *from our existing sample*, and **nonparametric** means that we will

do this in a way that obviates assumptions about the shape of the null distribution—in contrast to **parametric** approaches that do rely on such assumptions). These techniques are sometimes called “bootstrapping” techniques.

The idea is, if the treatment truly had no effect on the outcome, then the observations would be **exchangeable** between the treatment and control groups. That is, there would not be systematic differences between the treatment and control groups. This property may or may not be true in our observed sample (after all, that’s why we’re doing a hypothesis test in the first place), but we can draw new samples from our existing sample in a manner that forces exchangeability.

To perform this kind of test with our tea-tasting data, we would randomly shuffle the ratings in our dataset while leaving the condition assignments fixed. If we did this thousands of times and computed the treatment effect in each case, the result would be a null distribution: what we might expect the treatment effect to look like if there was *no* condition effect. In essence, we’re using a simulated version of “random assignment” here to *break* the dependency between the condition manipulation and the observed data.

We can then compare our *actual* treatment effect to this nonparametric null distribution. If the actual treatment was smaller than the 2.5th percentile or larger than the 97.5th percentile in the null distribution, we would reject the null with  $p < 0.05$ , just the same as if we had used a

*t*-test.

Resampling-based tests are extremely useful in a wide variety of cases. They can sometimes be less powerful than parametric approaches and they almost always require more computation, but their versatility makes them a great generic tool for data analysis.

2689

## 2690 *6.4 Inference and its discontents*

2691 In earlier sections of this chapter, we reviewed NHST and Bayesian ap-  
2692 proaches to inference. Now it's time to step back and think about some  
2693 of the ways that inference practices—especially those related to NHST—  
2694 have been problematic for psychology research. We'll begin with some  
2695 issues surrounding *p*-values and then give a specific accident report re-  
2696 lated to the process of “*p*-hacking” and some general philosophical dis-  
2697 cussion of how statistical testing relates to human reasoning.

### 2698 *6.4.1 Problems with the interpretation of *p*-values*

2699 *p*-values are basically likelihoods, in the sense we introduced in the pre-  
2700 vious chapter.<sup>16</sup> They are the likelihood of the data under the null  
2701 hypothesis! This likelihood is a critical number to know—for comput-  
2702 ing the Bayes Factor among other reasons. But it doesn't tell us a lot of

2703 things that we might like to know!

2704 For example,  $p$ -values don't tell us the probability of the data under a  
2705 specific alternative hypothesis that we might be interested in—that's the  
2706 posterior probability  $p(H_1|\text{data})$ . When our tea-tasting  $t$ -test yielded  
2707  $t(46) = 2.86$ ,  $p = 0.006$ , that  $p$  is *not* the probability of the null hypoth-  
2708 esis being true! And it's definitely not the probability of milk-first tea  
2709 being better.

2710 What can you conclude when  $p > 0.05$ ? According to the classical  
2711 logic of NHST, the answer is “nothing”! A failure to reject the null  
2712 hypothesis doesn't give you any additional evidence *for* the null. Even  
2713 if the probability of the data (or some more extreme data) under  $H_0$   
2714 is high, their probability might be just as high or higher under  $H_1$ .<sup>17</sup>  
2715 But many practicing researchers make this mistake. Aczel et al. (2018)  
2716 coded a sample of articles from 2015 and found that 72% of negative  
2717 statements were inconsistent with the logic of their statistical paradigm  
2718 of choice—most were cases where researchers said that an effect was not  
2719 present when they had simply failed to reject the null.

2720 These are not the only issues with  $p$ -values. In fact, people have so  
2721 much trouble understanding what  $p$ -values *do* say that there are whole  
2722 articles written about these misconceptions. Table 6.3 shows a set of  
2723 misconceptions documented and refuted by S. N. Goodman (2008).

<sup>17</sup> Of course, weighing these two against each other brings you back to the Bayes Factor.

2724 Let's take a look at just a few. Misconception 1 is that, if  $p = 0.05$ , the  
2725 null has a 5% chance of being true. This misconception is a result of  
2726 confusing  $p(H_0|\text{data})$  (the posterior) and  $p(\text{data}|H_0)$  (the likelihood—  
2727 also known as the  $p$ -value). Misconception 2—that  $p > 0.05$  allows us to  
2728 *accept* the null—also stems from this reversal of posterior and likelihood.  
2729 And misconception 3 is a misinterpretation of the  $p$ -value as an effect  
2730 size (which we learned about in the last chapter): a large effect is likely  
2731 to be clinically important, but with a large enough sample size, you can  
2732 get a small  $p$ -value even for a very small effect. We won't go through  
2733 all the misconceptions here, but we encourage you to challenge yourself  
2734 to work through them (as in the exercise below).

Table 6.3  
A “dirty dozen”  $p$ -value misconceptions. Adapted from S. N. Goodman (2008).

Misconception
1 “If $p = 0.05$ , the null hypothesis has only a 5% chance of being true.”
2 “A nonsignificant difference (e.g., $p \geq 0.05$ ) means there is no difference between groups.”
3 “A statistically significant finding is clinically important.”
4 “Studies with $p$ -values on opposite sides of 0.05 are conflicting.”
5 “Studies with the same $p$ -value provide the same evidence against the null hypothesis.”
6 “ $p = 0.05$ means that we have observed data that would occur only 5% of the time under the null hypothesis.”
7 “ $p = 0.05$ and $p \leq 0.05$ mean the same thing.”
8 “ $p$ -values are properly written as inequalities (e.g., ‘ $p \leq 0.02$ ’ when $p = .015$ ).”

---

### Misconception

---

- 9 “ $p = 0.05$  means that if you reject the null hypothesis, the probability of a false positive error is only 5%.”
  - 10 “With a  $p = 0.05$  threshold for significance, the chance of a false positive error will be 5%.”
  - 11 “You should use a one-sided  $p$ -value when you don’t care about a result in one direction, or a difference in that direction is impossible.”
  - 12 “A scientific conclusion or treatment policy should be based on whether or not the  $p$  value is significant.”
- 

<sup>2735</sup> Beyond these misconceptions, there’s another problem. The  $p$ -value is  
<sup>2736</sup> a probability of a certain set of events happening (corresponding to the  
<sup>2737</sup> observed data or any “more extreme” data—that is to say, data further  
<sup>2738</sup> from the null). Since  $p$ -values are probabilities, we can combine them  
<sup>2739</sup> together across different events. If we run a “null experiment”—an ex-  
<sup>2740</sup> periment where the true effect is zero—the probability of a dataset with  
<sup>2741</sup>  $p < 0.05$  is of course 0.05. But if we run two such experiments, we can  
<sup>2742</sup> get  $p < 0.05$  with probability 0.1. By the time we run 20 experiments,  
<sup>2743</sup> we have an 0.64 chance of getting a positive result.

<sup>2744</sup> It would obviously be a major mistake to run 20 experiments and then  
<sup>2745</sup> report only the positive ones (which, by design, are false positives) as  
<sup>2746</sup> though these still were “statistically significant.” The same thing applies  
<sup>2747</sup> to doing 20 different statistical tests within a single experiment. There

2748 are many statistical corrections that can be made to adjust for this prob-  
2749 lem, which is known as the problem of **multiple comparisons**.<sup>18</sup> But  
2750 the broader issue is one of transparency: unless you *know* what the  
2751 appropriate set of experiments or tests is, it's not possible to implement  
2752 one of these corrections!<sup>19</sup>

### ❖ ACCIDENT REPORT

#### *Do extraordinary claims require extraordinary evidence?*

In a blockbuster paper that may have inadvertently kicked off the replication crisis, Bem (2011) presented nine experiments he claimed provided evidence for precognition—that participants somehow had foreknowledge of the future. In the first of these experiments, Bem showed each of a group of 100 undergraduates 36 two-alternative forced choice trials in which they had to guess which of two locations on a screen would reveal a picture immediately before the picture was revealed. By chance, participants should choose the correct side 50% of the time of course. Bem found that, specifically for erotic pictures, participants' guesses were 53.1% correct. This rate of guessing was unexpected under the null hypothesis of chance guessing ( $p = 0.01$ ). Eight other studies with a total of more than 1,000 participants yielded apparently supportive evidence, with participants appearing to show a variety of psychological effects even before the stimuli were shown!

<sup>18</sup> The simplest and most versatile one, the Bonferroni correction, just divides 0.05 (or technically, whatever your threshold is) by the number of comparisons you are making. Using that correction, if you do 20 null experiments, you would have a 3% chance of a false positive.

<sup>19</sup> This issue is especially problematic with  $p$ -values because they are so often presented as an independent set of tests, but the problem of multiple comparisons comes up when you compute a lot of independent Bayes Factors as well. “Posterior hacking” via selective reporting of Bayes Factors is perfectly possible (Simonsohn 2014).

Based on this evidence, should we conclude that precognition exists?

Probably not. Wagenmakers et al. (2011) presented a critique of Bem's findings, arguing that (1) Bem's experiments were exploratory (not pre-registered) in nature, (2) that Bem's conclusions were *a priori* unlikely, and (3) that the level of statistical evidence from his experiments was quite low. We find each of these arguments alone compelling; together they present a knockdown case against Bem's interpretation.

First, we've already discussed the need to be skeptical about situations where experimenters have the opportunity for analytic flexibility in their choice of measures, manipulations, samples, and analyses. Flexibility leads to the possibility of cherry-picking those set of decisions from the “garden of forking paths” that lead to a positive outcome for the researcher's favored hypothesis (for more details, see chapter 11). And there is plenty of flexibility on display even in experiment 1 of Bem's paper. Although there were 100 participants in the study, they may have been combined post hoc from two distinct samples of 40 and 60, each of which saw different conditions. The 40 made guesses about the location of erotic, negative, and neutral pictures; the 60 saw erotic, positive non-romantic, and positive romantic pictures. The means of each of these conditions were presumably tested against chance (at least six comparisons, for a false positive rate of 0.26). Had positive romantic pictures been found significant, Bem certainly could have interpreted this finding the same way he interpreted the erotic ones.

Second, as we discussed, a *p*-value close to 0.05 does not necessarily pro-

vide strong evidence against the null hypothesis. Wagenmakers et al. computed the Bayes Factor for each of the experiments in Bem's paper and found that, in many cases, the amount of evidence for  $H_1$  was quite modest under a default Bayesian  $t$ -test. Experiment 1 was no exception: the BF was 1.64, giving only "anecdotal" support for the hypothesis of some nonzero effect, even before the multiple-comparisons problem mentioned above.

Finally, since precognition is not supported by any prior compelling scientific evidence (despite many attempts to obtain such evidence) and defies well-established physical laws, perhaps we should assign a low prior probability to Bem's  $H_1$ , a nonzero precognition effect. Taking a strong Bayesian position, Wagenmakers et al. suggest that we might do well to adopt a prior reflecting how unlikely precognition is, say  $p(H_1) = 10^{-20}$ . And if we adopt this prior, even a very well-designed, highly informative experiment (with a Bayes factor conveying substantial or even decisive evidence) would still lead to a very low posterior probability of precognition.

Wagenmakers et al. concluded that, rather than supporting precognition, the conclusion from Bem's paper should be psychologists should revise how they think about analyzing their data (and avoid  $p$ -hacking)!

2756 6.4.1 *Philosophical (and empirical) views of probability*

2757 Up until now, we've presented Bayesian and frequentist tools as two  
2758 different sets of computations. But, in fact, these different tools derive  
2759 from fundamentally different philosophical perspectives on what a prob-  
2760 ability even is. Very roughly, frequentist approaches tend to believe that  
2761 probabilities quantify the long-run frequencies of certain events. So, if  
2762 we say that some outcome of an event has probability 0.5, we're say-  
2763 ing that if that event happened thousands of times, the long-run fre-  
2764 quency of the outcome would be 50% of the total events. In contrast,  
2765 the Bayesian viewpoint doesn't depend on this sense that events could  
2766 be exactly repeated. Instead, the **subjective Bayesian** interpretation of  
2767 probability is that it quantifies a person's degree of belief in a particular  
2768 outcome.<sup>20</sup>

2769 You don't have to take sides in this deep philosophical debate about  
2770 what probability is. But it's helpful to know that people actually seem  
2771 to reason about the world in ways that are well described by the subjec-  
2772 tive Bayesian view of probability. Recent cognitive science research has  
2773 made a lot of headway in describing reasoning as a process of Bayesian  
2774 inference where probabilities describe degrees of belief in different hy-  
2775 potheses (for a textbook review of this approach, see **N. D. Goodman,**  
2776 **Tenenbaum, and The ProbMods Contributors 2016**). These hypothe-

<sup>20</sup> This is really a very rough description. If you're interested in learning more about this philosophical background, we recommend the Stanford Encyclopedia of Philosophy entry, "Interpretations of Probability" (<https://plato.stanford.edu/entries/probability-interpret>).

ses in turn are a lot like the theories we described in chapter 2: they  
describe the relationships between different abstract entities (Tenenbaum et al. 2011). You might think that scientists are different from  
laypeople in this regard, but one of the striking findings from research  
on probabilistic reasoning and judgment is that expertise doesn't matter  
that much. Statistically trained scientists—and even statisticians—make  
many of the same reasoning mistakes as their untrained students (Kahneman and Tversky 1979). Even children seem to reason intuitively in  
a way that looks a bit like Bayesian inference (Gopnik 2012).

These cognitive science findings help to explain some of the problems  
that people (scientists included) have in reasoning about  $p$ -values. If  
you are an intuitively Bayesian reasoner, the quantity that you're prob-  
ably tracking is how much you believe in your hypothesis (its posterior  
probability). So, many people treat the  $p$ -value as the posterior proba-  
bility of the null hypothesis.<sup>21</sup> That's exactly what fallacy 1 in table 6.3  
states—"If  $p = 0.05$ , the null hypothesis has only a 5% chance of being  
true." But this equivalence is incorrect! Written in math,  $p(\text{data}|H_0)$   
(the likelihood that lets us compute the  $p$ -value) is not the same thing  
as  $p(H_0|\text{data})$  (the posterior that we want). Pulling from our accident  
report above, even if the *probability of the observed ESP data given the null  
hypothesis* is low, that doesn't mean that the *probability of ESP* is high.

<sup>21</sup> Cohen (1994) is a great treatment of this issue.

2798 6.4.2 *What framework to use?*

2799 The problem with binary inferences is that they enable behaviors that  
2800 can introduce bias into the scientific ecosystem. By the logic of statis-  
2801 tical significance, either an experiment “worked” or it didn’t. Because  
2802 everyone would usually rather have an experiment that worked than  
2803 one that didn’t, inference criteria like  $p$ -values often become a target  
2804 for selection, as we discussed in chapter 3.<sup>22</sup>

2805 If you want to quantify evidence for or against a hypothesis, it’s worth  
2806 considering whether Bayes Factors address your question better than  $p$ -  
2807 values. In practice,  $p$ -values are hard to understand and many people  
2808 misuse them—though to be fair, BFs are misused plenty too. These  
2809 issues may be rooted in basic facts about how human beings reason about  
2810 probability.

2811 Despite the reasons to be worried about  $p$ -values, for many practicing  
2812 scientists (at least at time of writing), there is no one right answer about  
2813 whether to use them or not. Even if we’d like to be Bayesian all the  
2814 time, there are a number of obstacles. First, though new computational  
2815 tools make fitting Bayesian models and extracting Bayes Factors much  
2816 easier than before, it’s still on average quite a bit harder to fit a Bayesian  
2817 model than it is a frequentist one. Second, because Bayesian analyses are

<sup>22</sup> More generally, this pattern is probably an example of Goodhart’s law, which states that when a measure becomes a target, it ceases to be a good measure (Strathern 1997). Once the outcomes of statistical inference procedures become targets for publication, they are subject to selection biases— $p$ -hacking, for example—that make them less meaningful.

2818 less familiar, it may be an uphill battle to convince advisors, reviewers,  
2819 and funders to use them.

2820 As a group of authors, some of us are more Bayesian than frequentist,  
2821 while others are more frequentist than Bayesian—but all of us recog-  
2822 nize the need to move between statistical paradigms depending on the  
2823 problem we’re working on. Furthermore, a lot of the time we’re not  
2824 so worried about which paradigm we’re using. The paradigms are at  
2825 their most divergent when making binary inferences, and they often  
2826 look much more similar when they are used in the context of quantify-  
2827 ing measurement precision.

## 2828 6.5 Computing precision

2829 Our last section presented an argument against using  $p$ -values for mak-  
2830 ing *dichotomous* inferences. But we still want to move from what we  
2831 know about our own limited sample to some inference about the pop-  
2832 ulation. How should we do this?

### 2833 6.5.1 Confidence intervals

2834 One alternative to binary hypothesis testing is to ask about the precision  
2835 of our estimates, in particular how similar an estimate from a particu-

2836 lar sample is to the population parameter of interest. For example, how  
2837 close is our tea-tasting effect estimate to the true effect in the popu-  
2838 lation? We don't know what the true effect is, but our knowledge of  
2839 sampling distributions lets us make some guesses about how precise our  
2840 estimate is.

2841 The **confidence interval** is a convenient frequentist way to summarize  
2842 the variability of the sampling distribution—and hence how precise our  
2843 point estimate is. The confidence interval represents the range of possi-  
2844 ble values for the parameter of interest that are plausible given the data.

2845 More formally, a 95% confidence interval for some estimate (call it  $\hat{\beta}$ , as  
2846 in our example) is defined as a range of possible values for  $\beta$  such that,  
2847 if we did repeated sampling, 95% of the intervals generated by those  
2848 samples would contain the true parameter,  $\beta$ .

2849 Confidence intervals (CIs) are constructed by estimating the middle 95%  
2850 of the sampling distribution of  $\hat{\beta}$ . Because of our hero, the central limit  
2851 theorem, we can treat the sampling distribution as normal for reasonably  
2852 large samples. Given this, it's common to construct a 95% confidence  
2853 interval  $\hat{\beta} \pm 1.96 \widehat{SE}$ .<sup>23</sup> If we were to conduct the experiment 100  
2854 times and calculate a confidence interval each time, we should expect  
2855 95 of the intervals to contain the true  $\beta$ , whereas we would expect the  
2856 remaining 5 to not contain  $\beta$ .<sup>24</sup>

<sup>23</sup> This type of CI is called a “Wald” confidence interval.

<sup>24</sup> In case you don't have enough tea to do the experiment 100 times to confirm this, you can do it virtually using this nice simulation tool: <https://istats.shinyapps.io/ExploreCoverage>.

2857 Confidence intervals are like betting on the inferences drawn from your  
2858 sample. The sample you drew is like one pull of a slot machine that will  
2859 pay off (i.e., have the confidence interval contain the true parameter)  
2860 95% of the time. Put more concisely: 95% of 95% confidence intervals  
2861 contain the true value of the population parameter.

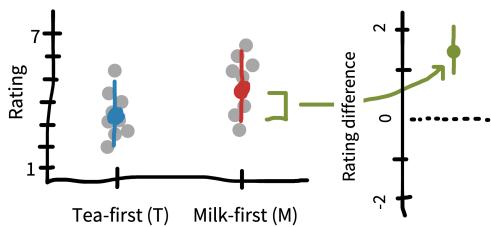


Figure 6.11  
Confidence intervals on each of the two condition estimates, as well as on the difference between conditions.

### CODE

Computing confidence intervals analytically is pretty easy. Here we first compute the standard error for the difference between conditions. The only tricky bit here is that we need to compute a pooled standard deviation.

```
tea_ratings <- filter(tea_data, condition == "tea first")$rating  
milk_ratings <- filter(tea_data, condition == "milk first")$rating  
  
n_tea <- length(tea_ratings)  
n_milk <- length(milk_ratings)  
sd_tea <- sd(tea_ratings)  
sd_milk <- sd(milk_ratings)  
  
tea_sd_pooled <- sqrt(((n_tea - 1) * sd_tea ^ 2 + (n_milk - 1) * sd_milk ^ 2) /  
(n_tea + n_milk - 2))  
  
tea_se <- tea_sd_pooled * sqrt((1 / n_tea) + (1 / n_milk))
```

Once we have the standard error, we can get the estimated difference between conditions and compute the confidence intervals by multiplying the standard error by 1.96.

```
delta_hat <- mean(milk_ratings) - mean(tea_ratings)  
tea_ci_lower <- delta_hat - tea_se * qnorm(0.975)  
tea_ci_upper <- delta_hat + tea_se * qnorm(0.975)
```

2863

2864 For visualization purposes, we can show the confidence intervals on  
2865 individual estimates (left side of figure 6.11). These tell us about the  
2866 precision of our estimates of each quantity relative to the population  
2867 estimate. But we've been talking primarily about the CI on the treat-

2868   ment effect  $\hat{\beta}$  (right side of figure 6.11). This CI allows us to make an  
2869   inference about whether or not it overlaps with zero—which is actually  
2870   equivalent in this case to whether or not the *t*-test is statistically signifi-  
2871   cant.

2872   6.5.2 *Confidence in confidence intervals?*

2873   Confidence intervals are often misinterpreted by students and re-  
2874   searchers alike (Hoekstra et al. 2014). Imagine a researcher conducts  
2875   an experiment and reports that “the 95% confidence interval for the  
2876   mean ranges from 0.1 to 0.4.” All of the statements in table 6.4, though  
2877   tempting to make about this situation, are *technically false*.

Table 6.4  
Confidence interval misconceptions for a confidence interval [0.1,0.4]. Adapted from  
Hoekstra et al. (2014).

Misconception
1   “The probability that the true mean is greater than 0 is at least 95%.”
2   “The probability that the true mean equals 0 is smaller than 5%.”
3   “The ‘null hypothesis’ that the true mean equals 0 is likely to be incorrect.”
4   “There is a 95% probability that the true mean lies between 0.1 and 0.4.”
5   “We can be 95% confident that the true mean lies between 0.1 and 0.4.”
6   “If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.”

2878   The problem with all of these statements is that, in the frequentist frame-

work, there is only one true value of the population parameter, and the variability captured in confidence intervals is about the *samples*, not the parameter itself.<sup>25</sup> For this reason, we can't make any statements about the probability of the value of the parameter or of our confidence in specific numbers. To reiterate, what we *can* say is: if we were to repeat the procedure of conducting the experiment and calculating a confidence interval many times, in the long run, 95% of those confidence intervals would contain the true parameter.

The Bayesian analog to a confidence interval is a **credible interval**. Recall that for Bayesians, parameters themselves are considered probabilistic (i.e., subject to random variation), not fixed. A 95% credible interval for an estimate,  $\hat{\beta}$ , represents a range of possible values for  $\beta$  such that there is a 95% probability that  $\beta$  falls inside the interval. Because we are now wearing our Bayesian hats, we are “allowed” to talk about  $\beta$  as if it were probabilistic rather than fixed. In practice, credible intervals are constructed by finding the posterior distribution of  $\beta$ , as in chapter 5, and then taking the middle 95%, for example.

Credible intervals are nice because they don't give rise to many of the inference fallacies surrounding confidence intervals. They actually *do* represent our beliefs about where  $\beta$  is likely to be, for example. Despite the technical differences between credible intervals and confidence in-

<sup>25</sup> In contrast, Bayesians think of parameters themselves as variable rather than fixed.

2900 tervals, in practice—with larger sample sizes and weaker priors—they  
2901 turn out to be quite similar to each other in many cases.<sup>26</sup>

## 2902 6.6 Chapter summary: Inference

2903 Inference tools help you move from characteristics of the sample to char-  
2904 acteristics of the population. This move is a critical part of generaliza-  
2905 tion from research data. But we hope we've convinced you that infer-  
2906 ence doesn't have to mean making a binary decision about the presence  
2907 or absence of an effect. A strategy that seeks to estimate an effect and  
2908 its associated precision is often much more helpful as a building block  
2909 for theory. As we move toward estimating causal effects in more com-  
2910 plex experimental designs, the process will require more sophisticated  
2911 models. Toward that goal, the next chapter provides some guidance for  
2912 how to build such models.

<sup>26</sup> They can diverge sharply in cases with less data or stronger priors (Morey et al. 2016), but in our experience this is relatively rare.



### DISCUSSION QUESTIONS

1. Can you write the definition of a  $p$ -value and a Bayes Factor without looking them up? Try this out—what parts of the definitions did you get wrong?
2. Take three of Goodman's (2008) "dirty dozen" in table 6.3) and write a description of why each is a misconception. (These can be checked against the original article, which gives a nice discussion of each.)



## READINGS

- Many of the concepts described here are illustrated beautifully via interactive visualizations. We recommend <https://seeing-theory.brown.edu> for a broad overview of statistical concepts and <https://rpsychologist.com/viz> for a number of interactives that specifically illustrate concepts discussed in this chapter and the previous one, including  $p$ -values, effect sizes, maximum likelihood estimation, confidence intervals, and Bayesian inference.
- A fun, polemical critique of NHST: Cohen, Jacob (1994). “The Earth is Round ( $p < .05$ ).” *American Psychologist* 49 (12):997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>.
- A nice introduction to Bayesian data analysis: Kruschke, John K., and Torrin M. Liddell (2018a). “Bayesian Data Analysis for Newcomers.” *Psychonomic Bulletin & Review* 25 (1): 155–77. <https://doi.org/10.3758/s13423-017-1272-1>.

2914

<sup>2915</sup> *References*

- Aczel, Balazs, Bence Palfi, Aba Szollosi, Marton Kovacs, Barnabas Szaszi, Peter Szecsi, Mark Zrubka, Quentin F. Gronau, Don van den Bergh, and Eric-Jan Wagenmakers. 2018. “Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation.” *Advances in Methods and Practices in Psychological Science* 1 (3): 357–366.

2916

- Bem, Daryl J. 2011. “Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect.” *Journal of Personality and Social Psychology* 100 (3): 407.
- Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2018. “Redefine Statistical Significance.” *Nature Human Behaviour* 2 (1): 6–10. <https://doi.org/10.1038/s41562-017-0189-z>.
- Cohen, Jacob. 1990. “Things I Have Learned (So Far).” *American Psychologist* 45:1304–1312.
- Cohen, Jacob. 1994. “The Earth is Round ( $p < .05$ ).” *American Psychologist* 49 (12): 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>.
- Cumming, Geoff. 2014. “The New Statistics: Why and How.” *Psychological Science* 25 (1): 7–29.
- Fisher, Ronald A. 1949. *The Design of Experiments*. 5th ed. Oliver & Boyd.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. Chapman; Hall/CRC.
- Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press.
- Gigerenzer, Gerd. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. 12. Cambridge University Press.
- Goodman, Noah D., Joshua B. Tenenbaum, and The ProbMods Contributors. 2016. “Probabilistic Models of Cognition.” <http://probmods.org/>.
- 2917 Goodman, Steven N. 1999. “Toward Evidence-Based Medical Statistics. 2:

- The Bayes Factor.” *Annals of Internal Medicine* 130 (12): 1005–1013.
- Goodman, Steven N. 2008. “A Dirty Dozen: Twelve p-Value Misconceptions.” In *Seminars in Hematology*, 45:135–140. 3. Elsevier.
- Gopnik, Alison. 2012. “Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications.” *Science* 337 (6102): 1623–1627.
- Hoekstra, Rink, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. “Robust Misinterpretation of Confidence Intervals.” *Psychonomic Bulletin & Review* 21 (5): 1157–1164.
- Jeffreys, Harold. 1961. *The Theory of Probability*. 3rd ed. OUP Oxford.
- Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision Under Risk.” *Econometrica* 47 (2): 363–391.
- Kruschke, John K. 2014. *Doing Bayesian Data Analysis: A Tutorial with r, JAGS, and Stan*. Academic Press.
- Kruschke, John K., and Torrin M. Liddell. 2018a. “Bayesian Data Analysis for Newcomers.” *Psychonomic Bulletin & Review* 25 (1): 155–177.
- Kruschke, John K., and Torrin M. Liddell. 2018b. “The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Power Analysis from a Bayesian Perspective.” *Psychonomic Bulletin & Review* 25 (1): 178–206.
- McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.
- Morey, Richard D., Rink Hoekstra, Jeffrey N. Rouder, Michael D. Lee, and Eric-Jan Wagenmakers. 2016. “The Fallacy of Placing Confidence in Confidence Intervals.” *Psychonomic Bulletin & Review* 23 (1): 103–123.

Morey, Richard D., and Jeffrey N. Rouder. 2011. “Bayes Factor Approaches for Testing Interval Null Hypotheses.” *Psychological Methods* 16 (4): 406.

Morey, Richard D., and Jeffrey N. Rouder. 2023. *BayesFactor: Computation of Bayes Factors for Common Designs*. <https://CRAN.R-project.org/package=BayesFactor>.

Pratt, John Winsor, Howard Raiffa, Robert Schlaifer, et al. 1995. *Introduction to Statistical Decision Theory*. MIT Press.

Simonsohn, Uri. 2014. “Posterior-Hacking: Selective Reporting Invalidates Bayesian Results Also.” <https://doi.org/10.2139/ssrn.2374040>.

Strathern, Marilyn. 1997. “‘Improving Ratings’: Audit in the British University System.” *European Review* 5 (3): 305–321.

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. “How to Grow a Mind: Statistics, Structure, and Abstraction.” *Science* 331 (6022): 1279–1285.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, and Han L. J. Van Der Maas. 2011. “Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011).” *Journal of Personality and Social Psychology* 100 (3): 426–432. <https://doi.org/10.1037/a0022790>.

## <sup>2920</sup> 7 MODELS



### LEARNING GOALS

- Articulate a strategy for estimating experimental effects using statistical models
- Build intuitions about how classical statistical tests relate to linear regression models
- Explore variations of the linear model, including generalized linear models and mixed effects models
- Reason about trade-offs and strategies for model specification, including the use of control variables

<sup>2921</sup>

<sup>2922</sup> In the previous two chapters, we introduced concepts surrounding es-  
<sup>2923</sup> timation of an experimental effect and inference about its relationship  
<sup>2924</sup> to the effect in the population. The tools we introduced there are for  
<sup>2925</sup> fairly specific research questions, and so are limited in their applicabil-  
<sup>2926</sup> ity. Once you get beyond the world of two-condition experiments in

2927 which each participant contributes one data point from a continuous  
2928 measure, the simple *t*-test is not sufficient.

2929 In some statistics textbooks, the next step would be to present a whole  
2930 host of other statistical tests that are designed for other special cases. We

2931 could even show a decision-tree: You have repeated measures? Use Test  
2932 X! Or categorical data? Use Test Y! Or three conditions? Use Test Z!

2933 But this isn't a statistics book, and even if it were, we don't advocate that  
2934 approach. The idea of finding a specific narrowly tailored test for your

2935 situation is part and parcel of the dichotomous NHST approach that  
2936 we tried to talk you out of in the last chapter. If all you want is your

2937  $p < 0.05$ , then it makes sense to look up the test that can allow you  
2938 to compute a  $p$  value in your specific case. But we prefer an approach

2939 that is more focused on getting a good estimate of the magnitude of the  
2940 causal effect.

2941 In this chapter, we begin to explore how to select an appropriate statis-  
2942 tical model to clearly and flexibly reason about these effects. A statistical

2943 model is a way of writing down a set of assumptions about how partic-  
2944 ular data are generated, the **data generating process**. Statistical models

2945 are the bread and butter tools for estimating particular **parameters** of  
2946 interest from empirical data—like the magnitude of a causal effect as-

2947 sociated with an experimental manipulation. They can also quantify

2948 MEASUREMENT PRECISION.

2949 For example, suppose you watch someone tossing a coin and observe a  
2950 sequence of heads and tails. A simple statistical model might assume that  
2951 the observed data are generated via the flip of a weighted coin. From the  
2952 perspective of the last two chapters, we could estimate a standard error  
2953 for the estimated proportion of flips that are heads (e.g., for six heads  
2954 out of eight flips, we have  $\hat{p} = 0.75 \pm 0.17$ ), or we could compare the  
2955 observed proportion against a null hypothesis. From a model-based per-  
2956 spective, however, we instead begin by thinking about where the data  
2957 came from: we might assume the coin being flipped has some weight  
2958 (a *latent*, or unobservable, parameter of the data generating process),  
2959 and our goal is to determine the most likely value of that weight given  
2960 the observed data. This single unified model can then also be used to  
2961 make inferences about whether the coin's weight differs from some null  
2962 model (a fair coin, perhaps), or to predict future flips.

2963 This example sounds a lot like the kinds of simple inferential tests we  
2964 talked about in the previous chapter; not very “model-y.” But things  
2965 get more interesting when there are multiple parameters to be estimated,  
2966 as in many real-world experiments. In the tea-tasting scenario we’ve  
2967 belabored over the past two chapters, a real experiment might involve  
2968 multiple people tasting different types of tea in different orders, all with

2969 some cups randomly assigned to be milk-first or tea-first. What we'll  
2970 learn to do in this chapter is to make a model of this situation that allows  
2971 us to reason about the magnitude of the milk-order effect while also  
2972 estimating variation due to different people, orders, and tea types. This  
2973 is the advantage of using models: once you are able to reason about  
2974 estimation and inference in model-based terms, you will be set free from  
2975 long decision trees and will be able to flexibly make the assumptions that  
2976 make sense for your data.<sup>1</sup>

2977 We'll begin by discussing the ubiquitous framework for building statisti-  
2978 cal models, **linear regression**.<sup>2</sup> We will then build connections between  
2979 regression and the *t*-test. This section will discuss how to add covariates  
2980 to regression models, and when linear regression does and doesn't work.  
2981 In the following section, we'll discuss the **generalized linear model**, an  
2982 innovation that allows us to make models of a broader range of data  
2983 types, including **logistic regression**. We'll then briefly introduce **mixed**  
2984 **models**, which allow us to model clustering in our datasets (such as clus-  
2985 ters of observations from a single individual or single stimulus item).  
2986 We'll end with some opinionated practical advice on model building.  
  
2987 If you're interested in building up intuitions about statistical model  
2988 building, then we recommend reading this chapter all the way through.  
2989 On the other hand, if you are already engaged in data analysis and

<sup>1</sup> We won't explore the connection to DAGs and Bayesian models here, but one way to think of this model building is as creating a causal theory of the experiment. This approach, which is advocated by McElreath (2018), creates powerful connections between the ideas about theory we presented in chapters 1 and 2 and the ideas about models here. If this sounds intriguing, we encourage you to go down the rabbit hole!

<sup>2</sup> The name regression originally comes from Galton's (1877) work on heredity. He was looking at the relationship between the heights of parents and children. He found that children's heights regressed, and he did so by creating a *regression model*. Now we use the term "regression" to mean any model of this form.

2990 want to see an example, we suggest that you skip to the last section,  
2991 where we give some opinionated practical advice on model building  
2992 and provide a worked example of fitting a mixed effects model and  
2993 interpreting it in context.

## 2994 7.1 Regression models

2995 There are many types of statistical models, but this chapter will focus  
2996 primarily on regression, a broad and extremely flexible class of mod-  
2997 els. A regression model relates a dependent variable to one or more  
2998 independent variables. Dependent variables are sometimes called **out-**  
2999 **come variables**, and independent variables are sometimes called **predic-**  
3000 **tor variables, covariates, or features**.<sup>3</sup> We will see that many common  
3001 statistical estimators (like the sample mean) and methods of inference  
3002 (like the *t*-test) are actually simple regression models. Understanding  
3003 this point will help you see many statistical methods as special cases of  
3004 the same underlying framework, rather than as unrelated, ad hoc tests.

### 3005 7.1.1 Regression for estimating a simple treatment effect

3006 Let's start with one of these special cases, namely estimating a treatment  
3007 effect,  $\beta$ , in a two-group design. In chapter 5, we solved this exact  
3008 challenge for the tea-tasting experiment. We applied a classical model

<sup>3</sup> The reverse is not true—not every predictor or covariate is an independent variable! One of the tricky things about relating regression models to causal hypotheses is that just because something is on the right side of a regression equation doesn't mean it's a causal manipulation. And of course, just because you've got an estimate of some predictor in a regression, that doesn't mean the estimate tells you about the magnitude of the *causal* effect. It could, but it also might not!

3009 in which the milk-first ratings are assumed to be normally distributed  
 3010 with mean  $\theta_M = \theta_T + \beta$  and standard deviation  $\sigma$ .<sup>4</sup>

3011 Let's now write that model as a regression model—that is, as a model  
 3012 that predicts each participant's tea rating,  $Y_i$ , given that participant's  
 3013 treatment assignment,  $X_i$ .  $X_i = 0$  represents the control (milk-first)  
 3014 group and  $X_i = 1$  represents the treatment (tea-first) group.<sup>5</sup> Here,  $Y_i$   
 3015 is the dependent variable and  $X_i$  is the independent variable. The sub-  
 3016 scripts  $i$  index the participants. To make this concrete, you can see some  
 3017 sample tea-tasting data (the first three observations from each condition)  
 3018 below (table 7.1), with the index  $i$ , the condition and its predictor  $X_i$ ,  
 3019 and the rating  $Y$ .

3020 Let's write this model more formally as a **linear regression of Y on X**.  
 3021 Conventionally, regression models are written with  $\beta$  symbols for all pa-  
 3022 rameters, so we'll now use  $\beta_0 = \theta_M$  for the mean in the milk-first group  
 3023 and  $\beta_1 = \theta_T - \theta_M$  as the average difference between the tea-first and  
 3024 milk-first groups. This  $\beta$  is a generalization of the one we were using  
 3025 to denote the causal effect above and in the previous two chapters.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

<sup>4</sup> Here's a quick reminder that "model" here is a way of saying "set of assumptions about the data generating procedure." So saying that some equation is a "model" is the same as saying that we think this is where the data came from. We can "turn the crank"—that is, generate data through the process that's specified in those equations, such as by pulling numbers from a normal distribution with mean  $\theta_T + \beta$  and standard deviation  $\sigma$ . In essence, we're committing to the idea that this process will give us data that are substantively similar to the ones we have already.

<sup>5</sup> Using 0 and 1 is known as **dummy coding** and allows us to interpret the parameter as the difference of the treatment group (tea-first) from the baseline (milk-first). There are many other ways to code categorical variables, with other interpretations. As a practical tip, be careful to check how your variables are coded before reporting anything!

3026 The term  $\beta_0 + \beta_1 X_i$  is called the **linear predictor**, and it describes the  
 3027 expected value of an individual's tea rating,  $Y_i$ , given that participant's  
 3028 treatment group  $X_i$  (the single independent variable in this model).  
 3029 That is, for a participant in the control group ( $X_i = 0$ ), the linear  
 3030 predictor is just equal to  $\beta_0$ , which is indeed the mean for the control  
 3031 group that we specified above. On the other hand, for a participant in  
 3032 the treatment group, the linear predictor is equal to  $\beta_0 + \beta_1$ , which  
 3033 is the mean for the treatment group that we specified. In regression  
 3034 jargon,  $\beta_0$  and  $\beta_1$  are **regression coefficients**, where  $\beta_1$  represents the  
 3035 association of the independent variable  $X$  with the outcome  $Y$ .

3036 The term  $\epsilon_i$  is the **error term**, referring to random variation of partici-  
 3037 pants' ratings around the group mean.<sup>6</sup> Note that this is a very specific  
 3038 kind of "error"; it does not include "error" due to bias, for example.  
 3039 Instead, you can think of the error terms as capturing the "error" that  
 3040 would be associated with predicting any given participant's rating based  
 3041 on just the linear predictor. If you predicted a control group partici-  
 3042 pant's rating as  $\beta_0$ , that would be a good guess—but you still expect the  
 3043 participant's rating to deviate somewhat from  $\beta_0$  (i.e., due to variability  
 3044 across participants beyond what is captured by their treatment groups).  
 3045 In our regression model, the linear predictor and error terms together  
 3046 say that participants' ratings scatter randomly (in fact, normally) around

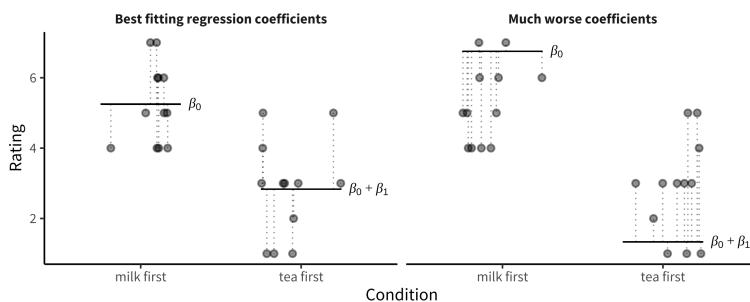
Table 7.1  
Example tea tasting data.

id	condition	X	rating (Y)
1	milk first	0	6
2	milk first	0	4
3	milk first	0	5
4	tea first	1	1
5	tea first	1	3
6	tea first	1	5

<sup>6</sup> Formally, we'd write  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ . The tilde means "is distributed as," and what follows is a normal distribution with mean 0 and variance  $\sigma^2$ .

3047 their group means with standard deviation  $\sigma$ . And that is exactly the  
 3048 same model we posited in chapter 5.<sup>7</sup>

3049 Now we have the model. But how do we estimate the regression co-  
 3050 efficients  $\beta_0$  and  $\beta_1$ ? The usual method is called **ordinary least squares**  
 3051 (**OLS**). Here's the basic idea. For any given regression coefficient es-  
 3052 timates  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , we would obtain different **predicted values**,  $\hat{Y}_i =$   
 3053  $\hat{\beta}_0 + \hat{\beta}_1 X_i$ , for each participant. Some regression coefficient estimates  
 3054 will yield better predictions than others. Ordinary least squares es-  
 3055 timation is designed to find the values of the regression coefficients that  
 3056 optimize these predictions, meaning that the predictions are as close as  
 3057 possible to participants' true outcomes,  $Y_i$ .



3058 Figure 7.1 illustrates the tea-tasting data for each condition (the dots)  
 3059 along with the model predictions for each condition  $\beta_0$  and  $\beta_0 + \beta_1$  (blue  
 3060 lines). The gap between each data point and the corresponding predic-  
 3061 tions (the thing that OLS wants to minimize) is shown by the dotted  
 3062 lines.<sup>8</sup> These distances are sample estimates, called **residuals**, of the true  
 3063 errors ( $\epsilon_i$ ). The left-hand plot shows the OLS coefficient values—the

<sup>7</sup> You may be wondering why so much effort was put into building boutique solutions for these special cases when a unified framework was available the whole time. A partial answer is that the classical infrastructure of statistics was developed before computers were widespread, and these special cases were chosen because they were easy to work with “analytically” (meaning to work out all the math by hand, using values from big numerical tables). Now that we have computers with more flexible algorithms, the model-based perspective is more practi-

Figure 7.1 accessible than it used to be.  
 (left) Best-fitting regression coefficients for the tea-tasting experiment. (right) Much worse coefficients for the same data. Dotted lines: residuals. Circles: data points for individual participants.

3064 ones that move the model's predictions as close as possible to the data  
 3065 points, in the sense of minimizing the total squared length of the dashed  
 3066 lines. The right-hand plot shows a substantially worse set of coefficient  
 3067 values.

3068 You'll notice that we aren't talking much about *p*-values in this chapter.  
 3069 Regression models can be used to produce *p*-values for specific coef-  
 3070 ficients, representing inferences about the likelihood of the observed  
 3071 data under some null hypothesis regarding the coefficients. You can  
 3072 also compute Bayes Factors for specific regression coefficients, or use  
 3073 Bayesian inference to fit these coefficients under some prior expecta-  
 3074 tion about their distribution. We won't talk much about this, or more  
 3075 generally how to fit the models we describe. As we said, we're not going  
 3076 to give a full treatment of all the relevant statistical topics. Instead we  
 3077 want to help you begin thinking about making models of your data.

<sup>8</sup> Ordinary least squares minimizes squared error loss, in the sense that it will choose the regression coefficient estimates whose predictions minimize  $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ , where  $n$  is the sample size. A wonderful thing about OLS is that those optimal regression coefficients (generically termed  $\hat{\beta}$ ) turn out to have a very simple closed-form solution:  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . We are using more general notation here that supports multiple independent variables:  $\hat{\beta}$  is a vector,  $\mathbf{X}$  is a matrix of independent variables for each subject, and  $\mathbf{y}$  is a vector of participants' outcomes. As more good news, the standard error for  $\hat{\beta}$  has a similarly simple closed form!

## CODE

As it turns out, fitting an OLS regression model in R is extremely easy. The underlying function is `lm()`, which stands for linear model. You can fit the model with a single call to this function with a “formula” as its argument. Here's the call:

```
mod <- lm(rating ~ condition, data = tea_data)
```

Formulas in R are a special kind of terse notation for regression equations where you write the outcome,  $\sim$  (distributed as), and the predictors. R assumes that you want an intercept by default, and there are also a number of other handy defaults that make R formulas a nice easy way to specify relatively complex regression models, as we'll see below.

Once you've fit the model and assigned it to a variable, you can call `summary()` to see a summary of the parameters of the model:

```
summary(mod)
```

You can also extract the coefficient values using `coef(mod)` and put them in a handy dataframe using `tidy(mod)` from the `broom` package (Robinson, Hayes, and Couch 2023).

3079

### 3080 7.1.2 Adding predictors

3081 The regression model we just wrote down is the same model that un-  
3082 derlies the *t*-test from chapter 6. But the beauty of regression modeling  
3083 is that much more complex estimation problems can also be written as  
3084 regression models that extend the model we made above. For example,  
3085 we might want to add another predictor variable, such as the age of the  
3086 participant.<sup>9</sup>

3087 Let's add this new independent variable and a corresponding regression

3088 coefficient to our model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

3089 Now that we have multiple independent variables, we've labeled them

3090  $X_1$  (treatment group) and  $X_2$  (age) for clarity.

3091 To illustrate how to interpret the regression coefficients in this model,

3092 let's use the linear predictor to compare the model's predicted tea ratings

3093 for two hypothetical participants who are both in the treatment group:

3094 20-year-old Alice and 21-year-old Bob. Alice's linear predictor tells us

3095 that her expected rating is  $\beta_0 + \beta_1 + \beta_2 \cdot 20$ . In contrast, Bob's linear

3096 predictor is  $\beta_0 + \beta_1 + \beta_2 \cdot 21$ . We could therefore calculate the expected

3097 difference in ratings for 21-year-olds versus 20-year-olds by subtracting

3098 Alice's linear predictor from Bob's, yielding just  $\beta_2$ .

3099 We would get the same result if Alice and Bob were instead 50 and 51

3100 years old, respectively. This equivalence illustrates a key point about

3101 linear regression models in general:

3102 The regression coefficient represents the expected differ-

3103 ence in outcome when comparing any two participants

3104 who differ by 1 unit of the relevant independent variable,

3105 and who do not differ on any other independent variables

3106 in the model.

3107 Here, the coefficient compares participants who differ by one year of  
3108 age. In “Practical modeling considerations” below, we discuss whether  
3109 and when to “control for” additional variables (i.e., when to add them  
3110 to your model).

### 3111 7.1.3 *Interactions*

3112 In our running example, we now have two predictors: condition and  
3113 age. But what if the effect of condition varies depending on the age  
3114 of the participant? This situation would correspond to a case where  
3115 (say) older people were more sensitive to tea ordering, perhaps because  
3116 of their greater tea experience. We call this an **interaction** effect: the  
3117 effect of one predictor depends on the state of another.

3118 Interaction effects are easily accommodated in our modeling framework.  
3119 We simply add a term to our model that is the product of condition ( $X_1$ )  
3120 and age ( $X_2$ ), and weight this product by another beta, which represents  
3121 the strength of this interaction:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \epsilon_i$$

3122 Statistical interactions are a very powerful modeling tool that can help  
3123 us understand the relationship between different experimental manip-  
3124 ulations or between manipulationes and covariates (such as age). We

3125 discuss their role in experimental design—as well as some of the inter-  
3126 pretive challenges that they pose—in much more detail in chapter 9.<sup>10</sup>

3127 7.1.4 *When does linear regression work?*

3128 Linear regression modeling with OLS is an incredibly powerful tech-  
3129 nique for creating models to estimate the influence of multiple predic-  
3130 tors on a single dependent variable. In fact, OLS is in a mathematical  
3131 sense the *best* way to fit a linear model!<sup>11</sup> But OLS only “works”—in  
3132 the sense of yielding good estimates—if three big conditions are met.

3133 1. **The relationship between the predictor and outcome must be lin-**  
3134 **ear.** In our comparison of Alice’s and Bob’s expected outcomes  
3135 based on their one-year age difference, we were able to interpret  
3136 the coefficient  $\beta_2$  as the average difference in  $Y_i$  when comparing  
3137 participants who differ by one year of age, *regardless* of whether  
3138 those ages are 20 vs 21 or 50 vs 51. If we believed this relation-  
3139 ship was **nonlinear**, then we could transform our predictor—for  
3140 example, including a **quadratic** effect of age by adding a  $\beta_3 * X_2^2$   
3141 term. The *relationship* between this new predictor and the out-  
3142 come would still be linear, however. It is always a good idea to  
3143 use visualizations like scatter plots to look for possible problems

<sup>10</sup> We won’t go into this topic here, but we do want to provide a pointer to one of the most persistent challenges that comes up when you specify regression models with categorical predictors—and especially their interactions: how you “code” these categorical predictors. Above we created a “dummy” variable  $X$  that encoded milk-first tea as 0 and tea-first tea as 1. Dummy variables are very easy to think about, but in models with interactions, they can cause some problems. Because the interaction in our example model is a product of the dummy-coded condition variable and age, the interaction term  $\beta_3$  is interpreted as the effect of age *for the tea-first condition* ( $X = 1$ ) and hence the effect of age  $\beta_2$  is actually the effect of age *for the milk-first condition*. The way to deal with this issue is to use a different coding system, such as **contrast coding**. Davis (2010) gives a good tutorial on this tricky topic.

3144 with assuming a linear relationship between a predictor and your  
3145 outcome.

3146 **2. Errors must be independent.** In our example, observations in the  
3147 regression model (i.e., rows in the dataset) were sampled indepen-  
3148 dently: each participant was recruited independently to the study  
3149 and each performed a single trial. On the other hand, suppose we  
3150 have repeated-measures data in which we sample participants and  
3151 then obtain multiple measurements for each participant. Within  
3152 each participant, measurements would likely be correlated (per-  
3153 haps because participants differ on their general level of tea enjoy-  
3154 ment). This correlation can invalidate inferences from a model  
3155 that does not accommodate the correlation. We'll discuss this  
3156 problem in detail below.

3157 **3. Errors must be normally distributed and unrelated to the predic-**  
3158 **tor.** Imagine older people have very consistent tea-ordering pref-  
3159 erences while younger people do not. In that case, the models' er-  
3160 ror term would be less variable for older participants than younger  
3161 ones. This issue is called **heteroskedasticity**. It is a good idea to  
3162 plot each independent variable versus the residuals to see if the  
3163 residuals are more variable for certain values of the independent  
3164 variable than for others.

<sup>11</sup> There is a precise sense in which OLS gives the *very best* predictions we could ever get from any model that posits linear relationships between the independent variables and the outcome. That is, you can come up with any other linear, unbiased model you want, and yet if the assumptions of OLS are fulfilled, predictions from OLS will always be less noisy than those of your model. This is because of an elegant mathematical result called the Gauss-Markov Theorem.

<sup>3165</sup> If any of these three conditions are violated, it can undermine the esti-  
<sup>3166</sup> mates and inferences you draw from your model.

## <sup>3167</sup> 7.2 *Generalized linear models*

<sup>3168</sup> So far we have considered continuous outcome measures, like tea rat-  
<sup>3169</sup> ings. What if we instead had a binary outcome, such as whether a par-  
<sup>3170</sup> ticipant liked or didn't like the tea, or a count outcome, such as the  
<sup>3171</sup> number of cups a participant chose to drink? These and other noncon-  
<sup>3172</sup> tinuous outcomes often violate the assumptions of OLS, in particular  
<sup>3173</sup> because they often induce heteroskedastic errors.

<sup>3174</sup> Binary outcomes inherently produce heteroskedastic errors because the  
<sup>3175</sup> variance of a binary variable depends directly on the outcome probabil-  
<sup>3176</sup> ity. Errors will be more variable when the outcome probability is closer  
<sup>3177</sup> to 0.50, and much less variable for when the probability is closer to 0  
<sup>3178</sup> or 1.<sup>12</sup> This heteroskedasticity in turn means that inferences from the  
<sup>3179</sup> model (e.g., *p*-values) can be incorrect, sometimes just a little bit off but  
<sup>3180</sup> sometimes dramatically incorrect.<sup>13</sup>

<sup>3181</sup> Happily, **generalized linear models** (GLMs) are regression models closely  
<sup>3182</sup> related to OLS that can handle noncontinuous outcomes. These models  
<sup>3183</sup> are called “generalized” because OLS is one of many members of this

<sup>12</sup> Specifically, the variance of a bi-  
nary variable with probability  $p$  is sim-  
ply  $p(1 - p)$ , which is largest when  $p =$   
0.50.

<sup>13</sup> Ordinary least squares can also be  
used with binary outcomes, in which  
case the coefficients represent differ-  
ences in probabilities. However, the  
usual model-based standard errors will  
be incorrect.

3184 large class of models. To see the connection, let's first write an OLS  
3185 model more generally in terms of what it says about the expected value  
3186 of the outcome, which we notate as  $E[Y_i]$ :

$$E[Y_i] = \beta_0 + \sum_{j=1}^p \beta_j X_{ij}$$

3187 where  $p$  is the number of independent variables,  $\beta_0$  is the intercept, and  
3188  $\beta_j$  is the regression coefficient for the  $j^{th}$  independent variable. This  
3189 equation is just a math-y way of saying that you predict from a regres-  
3190 sion model by adding up each of the predictors' contributions to the  
3191 expected outcome ( $\beta_j X_{ij}$ ).

3192 The linear predictor of a GLM (i.e.,  $\beta_0 + \sum_{j=1}^p \beta_j X_{ij}$ ) looks exactly the  
3193 same as for OLS, but instead of modeling  $E[Y_i]$ , a GLM models some  
3194 **transformation**,  $g(\cdot)$ , of the expectation:

$$g(E[Y_i]) = \beta_0 + \sum_{j=1}^p \beta_j X_{ij}$$

3195 GLMs involve transforming the *expectation* of the outcome, not the out-  
3196 come itself! That is, in GLMs, we are not just taking the outcome vari-  
3197 able in our dataset and transforming it before fitting an OLS model, but  
3198 rather we are fitting a different model entirely, one that posits a fun-  
3199 damentally different relationship between the predictors and the ex-  
3200 pected outcomes. This transformation is called the **link function**. In  
3201 other words, to fit different kinds of outcomes, all we need to do is con-

3202 struct a standard linear model and then just transform its output via the  
 3203 appropriate link function.

3204 Perhaps the most common link function is the **logit** link, which is suit-  
 3205 able for binary data. This link function looks like this, where  $w$  is any  
 3206 probability that is strictly between 0 and 1:

$$g(w) = \log\left(\frac{w}{1-w}\right)$$

3207 The term  $w/(1-w)$  is called the **odds** and represents the probability of  
 3208 an event occurring divided by the probability of its not occurring. The  
 3209 resulting model is called **logistic regression** and looks like:

$$\text{logit}(E[Y_{it}]) = \log\left(\frac{E[Y_i]}{1 - E[Y_i]}\right) = \beta_0 + \sum_{j=1}^p \beta_j X_{ij}$$

3210 Exponentiating the coefficients (i.e.,  $e^\beta$ ) would yield **odds ratios**, which  
 3211 are the *multiplicative* increase in the odds of  $Y_i = 1$  that is associated  
 3212 with a one-unit increase in the relevant predictor variable.

3213 Figure 7.2 shows the way that a logistic regression model transforms a  
 3214 predictor ( $X$ ) into an outcome probability that is bounded at 0 and 1.  
 3215 Critically, although the predictor is still linear, the logit link means that  
 3216 the same change in  $X$  can result in a different change in the absolute  
 3217 probability of  $Y$  depending on where you are on the  $X$  scale. In this  
 3218 example, if you are in the middle of the predictor range, a one-unit

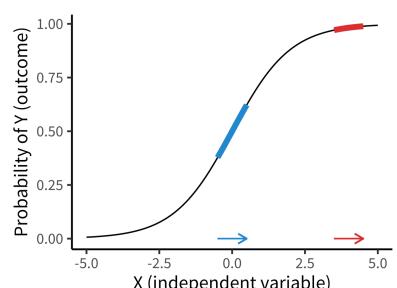


Figure 7.2

An example of how logistic regression transforms a change in the mean-centered predictor  $X$  into a change in the expected outcome  $Y$ . The same absolute change in  $X$  is associated in a large difference in the probability of the outcome when  $X$  is near its mean (blue) vs a small change in the outcome when  $X$  is large (red) or small.

3219 change in  $X$  results in a 0.24 change in probability (blue). At a higher  
3220 value, the change is much smaller (0.02). Notice how this is different  
3221 from the linear regression model above, where the same change in age  
3222 always resulted in the same change in preference!

### CODE

GLMs are as easy to fit in R as standard LMs. You simply need to call the `glm()` function—and to specify the link function. For our example above of a binary “liking” judgment, the call would be:

```
glm(liked_tea ~ condition, data = tea_data, family = "binomial")
```

The `family` argument specifies the type of distribution being used, where `binomial` is the logistic link function.

3223

3224 We have only scratched the surface of GLMs here. First, there are many  
3225 different link functions that are suitable for different outcome types.  
3226 And second, GLMs differ from OLS not only in their link functions  
3227 but also in how they handle the error terms. Our broader goal in this  
3228 chapter is to show you how regression models are *models of data*. In that  
3229 context, GLMs use link functions as a way to make models that generate  
3230 many different types of outcome data.<sup>14</sup>

<sup>14</sup> We sometimes think of linear models as a set of tinker toys you can snap together to stack up a set of predictors. In that context, link functions are an extra “attachment” that you can snap onto your linear model to make it generate a different response type.

### 3231 7.3 Linear mixed effects models

3232 Experimental data often contain multiple measurements for each par-  
3233 ticipant (so-called **repeated measures**). In addition, these measurements  
3234 are often based on a sample of stimulus items (which then each have mul-  
3235 tiple measures as well). This clustering is problematic for OLS models  
3236 because the error terms for each datapoint are not independent.

3237 Non-independence of datapoints may seem at first glance like a small  
3238 issue, but it can present a deep problem for making inferences. Take the  
3239 tea-tasting data we looked at above, where we had 24 observations in  
3240 each condition. If we fit an OLS model, we observe a highly significant  
3241 tea-first effect. Here is the estimate and confidence interval for that  
3242 coefficient:  $b = -2.42$ , 95% CI  $[-3.50, -1.33]$ . Based on what we  
3243 talked about in the previous chapter, it seems like we'd be licensed in  
3244 rejecting the null hypothesis that this effect is due to sampling variation  
3245 and interpret this instead as evidence for a generalizable difference in  
3246 tea preference in our sampled population.

3247 But suppose we told you that all of those 48 total observations (24 in  
3248 each condition) were from one individual named George. That would  
3249 change the picture considerably. Now we'd have no idea whether the  
3250 big effect we observed reflected a difference in the population, but we  
3251 would have a very good sense of what George's preference is!<sup>15</sup> The

<sup>15</sup> We discuss the strengths and weaknesses of repeated-measures designs like this in chapter 9 and the statistical trade-offs of having many people with a small number of observations per person vs a small number of people with many observations per person in chapter 10.

3252 confidence intervals and p-values from our OLS model would be wrong  
3253 now because all of the error terms would be highly correlated—they  
3254 would all reflect George’s preferences.

3255 How can we make models that deal with clustered data? There are  
3256 a number of widely used approaches for solving this problem including  
3257 **linear mixed effects models, generalized estimating equations, and clus-**  
3258 **tered standard errors** (often used in economics). Here we will illustrate  
3259 how the problem gets solved in linear mixed models, which are an ex-  
3260 tension of OLS models that are fast becoming a standard in many areas  
3261 of psychology (Bates et al. 2015).

### 3262 7.3.1 Modeling random variation in clusters

3263 In linear mixed effects models, we modify the linear predictor itself to  
3264 model differences across clusters. Instead of just measuring George’s  
3265 preferences, suppose we modified the original tea-tasting experiment  
3266 (without the age covariate) to collect ten ratings from each participant:  
3267 five milk-first and five tea-first. We define the model the same way as  
3268 we did before, with some minor differences:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_i + \epsilon_{it}$$

<sup>3269</sup> where  $Y_{it}$  is participant  $i$ 's rating in trial  $t$  and  $X_{it}$  is the participant's  
<sup>3270</sup> assigned treatment in trial  $t$  (i.e., milk-first or tea-first).

<sup>3271</sup> If you compare this equation to the OLS equation above, you will notice  
<sup>3272</sup> that we added two things. First, we've added subscripts that distinguish  
<sup>3273</sup> trials from participants. But the big one is that we added  $\gamma_i$ , a separate  
<sup>3274</sup> intercept value for each participant. We call this a **random intercept**  
<sup>3275</sup> because it varies across participants (who are randomly selected from  
<sup>3276</sup> the population).<sup>16</sup>

<sup>3277</sup> The random intercept means that we have assumed that each participant  
<sup>3278</sup> has their own typical "baseline" tea rating—some participants overall  
<sup>3279</sup> just like tea more than others—and these baseline ratings are normally  
<sup>3280</sup> distributed across participants. Thus, ratings are correlated within par-  
<sup>3281</sup> ticipants because ratings cluster around each participant's *distinct* base-  
<sup>3282</sup> line tea rating. This model is better able to block misleading inferences.

<sup>3283</sup> For example, suppose we only had one participant in each condition  
<sup>3284</sup> (say, George provided 24 milk-first ratings and Alice provided 24 tea-  
<sup>3285</sup> first ratings). If we found higher ratings in one condition, we would be  
<sup>3286</sup> able to attribute this difference to participant-level variation rather than  
<sup>3287</sup> to the treatment.<sup>17</sup>

<sup>3288</sup> Following the same logic, we could fit random intercepts for different  
<sup>3289</sup> stimulus items (for example, if we used different types of tea for dif-

<sup>16</sup> Formally, we'd note this random variation by saying that  $\gamma_i \sim N(0, \tau^2)$ —in other words, that participants' random intercepts are sampled from a normal distribution around the shared intercept  $\beta_0$  with standard deviation  $\tau$ .

<sup>17</sup> Of course, this would be a terrible experiment! Ideally, we would address this problem upstream in our experiment design; see chapter 9.

3290 ferent trials). We modeled participants as having normally distributed  
3291 variation, and we can model stimulus variation the same way. Each  
3292 stimulus item is assumed to produce a particular average outcome (i.e.,  
3293 some teas are tastier than others), with these average outcomes sampled  
3294 from a normally distributed population.

### CODE

Remarkably, GLMMs are not much harder to specify in R than standard  
LMs. One very popular package is `lme4` (?), which provides the `lmer()`  
and `glmer()` functions (the latter for generalized linear mixed effect mod-  
els). For our example here, we'd write:

```
library(lme4)  
  
lmer(rating ~ condition + (1 | id), data = tea_data)
```

In this model, the syntax `(1 | id)` specifies that we want a random in-  
tercept for each level of `id`.

3295

#### 7.3.2 Random slopes and the challenges of mixed effects models

3296 Linear mixed effects models can be further extended to model cluster-  
3297 ing of the independent variables' *effects* within subjects, not just clus-  
3298 tering of average *outcomes* within subjects. To do so, we can introduce  
3299 **random slopes** ( $\delta_i$ ) to the model, which are multiplied by the condition  
3300 variable  $X$  and represent differences across participants in the effect of  
3301

3302 tea tasting:

$$Y_i = \beta_0 + \beta_1 X_{it} + \gamma_i + \delta_i X_{it} + \epsilon_{it}$$

3303 Just like the random intercepts, these random slopes will be assumed to

3304 vary across participants, following a normal distribution.<sup>18</sup>

3305 This model now describes random variation in both overall how much

3306 someone likes tea *and* how strong their ordering preference is. Both of

3307 these likely do vary in the population, and so it seems like a good thing

3308 to put these in your model. Indeed, under some circumstances, adding

3309 random slopes is argued to be very important for making appropriate

3310 inferences.<sup>19</sup>

<sup>18</sup> These random slopes and intercepts can be assumed to be independent or correlated with one another, depending on the modeler's preference.

## CODE

Specifying random slopes in the `lme4` package is also relatively straightforward:

```
lmer(rating ~ condition + (condition | id), data = tea_data)
```

Here, `(condition | id)` means “a separate random slope for `condition` should be fit for each level of `id`.” Of course, specifying such a model is easier than fitting it correctly.

<sup>19</sup> There's lots of debate in the literature about the best random effect structure for mixed effects models. This is a very tricky and technical subject. In brief, some folks argue for so-called **maximal** models, in which you include every random effect that is justified by the design (Barr et al. 2013). Here that would mean including random slopes for each participant. The problem is that these models can get very complex and can be very hard to fit using standard software. We won't weigh in on this topic, but as you start to use these models on more complex experimental designs, it might be worth reading up on.

3311 On the other hand, the model is much more complicated. When we

3312 had a simple OLS model above, we had only two parameters to fit ( $\beta_0$

3313 and  $\beta_1$ ), but now we have those two plus two more, representing the

3315 standard deviations of the individual participant intercepts and slopes,  
3316 plus parameters for each participant and for the condition effect for each  
3317 participant. So we went from two parameters to  $24!$ <sup>20</sup> This complexity  
3318 can lead to problems in fitting the models, especially with very small  
3319 datasets (where these parameters are not very well constrained by the  
3320 data) or very large datasets (where computing all these parameters can  
3321 be tricky).<sup>21</sup>

3322 More generally, linear mixed effects models are very flexible, and they  
3323 have become quite common in psychology. But they do have signifi-  
3324 cant limitations. As we discussed, they can be tricky to fit in standard  
3325 software packages. Further, the accuracy of these models relies on our  
3326 ability to specify the structure of the random effects correctly.<sup>22</sup> If we  
3327 specify an incorrect model, our inferences will be wrong! But it is some-  
3328 times difficult to know how to check whether your model is reasonable,  
3329 especially with a small number of clusters or observations.

## 3330 7.4 How do you use models to analyze data?

3331 In the prior parts of this chapter, we've described a suite of regression-  
3332 based techniques—standard OLS, the generalized linear model, and lin-  
3333 ear mixed effects models—that can be used to model the data result-  
3334 ing from randomized experiments (as well as many other kinds of data).

<sup>20</sup> Though we should note that these pa-  
rameters aren't technically all indepen-  
dent from one another due to the struc-  
ture of the mixed effect model.

<sup>21</sup> Many R users may be familiar with  
the widely used `lme4` package for fit-  
ting mixed effects models using frequen-  
tist tools related to maximum likeli-  
hood. Such models can also be fit us-  
ing Bayesian inference with the `brms`  
package (Bürkner 2021), which provides  
many powerful methods for specifying  
complex models.

<sup>22</sup> One particularly problematic situa-  
tion is when the correlation structure  
of the errors is mis-specified, for exam-  
ple if observations within a participant  
are more correlated for participants in  
the treatment group than in the control  
group; in such cases, mixed model esti-  
mates can be substantially biased (Bie et  
al. 2021).

3335 The advantage of regression models over the simpler estimation and in-  
3336 ference methods we described in the prior two chapters is that these  
3337 models can more effectively take into account a range of different kinds  
3338 of variation, including covariates, multiple manipulations, and clustered  
3339 structure. Further, when used appropriately to analyze a well-designed  
3340 randomized experiment, regression models can give an unbiased esti-  
3341 mate of a causal effect of interest, our main goal in doing experiments.

3342 But—practically speaking—how should you about building a model  
3343 for your experiment? What covariates should you include, and what  
3344 should you leave out? There are many ways to use models to explore  
3345 datasets, but in this section we will try to sketch a default approach for  
3346 the use of models to estimate causal effects in experiments in the most  
3347 straightforward way. Think of this as a starting point. We'll begin this  
3348 section by giving a set of rules of thumb, then discuss a worked example.

3349 Our final subsections will deal with the issues of when you should in-  
3350 clude covariates in your model and how to check if your result is robust  
3351 across multiple different model specifications.

 DEPTH

## *An alternative approach: Generalized estimating equations*

A second class of methods that helps resolve issues of clustering is **generalized estimating equations** (GEE). In this approach, we leave the linear predictor alone. We do not add random intercepts or slopes, nor do we assume anything about the distribution of the errors (i.e., we no longer assume that they are normal, independent, and homoskedastic).

In GEE, we instead provide the model with an initial “guess” about how we think the errors might be related to one another; for example, in a repeated-measures experiment, we might guess that the errors are exchangeable, meaning that they are correlated to the same degree within each participant but are uncorrelated across participants. Instead of *assuming* that our guess is correct, as do linear mixed models (LMM), GEE estimates the correlation structure of the errors empirically, using our guess as a starting point, and it uses this correlation structure to arrive at point estimates and inference for the regression coefficients. Remarkably, as the number of clusters and observations become very large, GEE will *always* provide unbiased point estimates and valid inference, *even if* our guess about the correlation structure was bad. Additionally, with simple finite-sample corrections (Mancl and DeRouen 2001), GEE seems to provide valid inference at smaller numbers of clusters than does LMM.

The price paid for these nice safeguards against model misspecification

is that, in principle, GEE will typically have less statistical power than LMM *if* the LMM is in fact correctly specified, but the difference may be surprisingly slight in practice (Bie et al. 2021). For these reasons, some of this book’s authors actually favor GEE with finite-sample corrections over LMM as the default model for clustered data, though they are much less common in psychology.

3353

3354    7.4.1 *Modeling rules of thumb*

3355    Our approach to statistical modeling is to start with a “default model”  
3356    that is known in the literature as a **saturated model**. The saturated model  
3357    of an experiment includes the full design of the experiment—all main  
3358    effects and interactions—and nothing else. If you are manipulating a  
3359    variable, include it in your model. If you are manipulating two, in-  
3360    clude them both and their interaction. If your design includes repeated  
3361    measurements for participants, include a random effect of participant;  
3362    if it includes experimental items for which repeated measurements are  
3363    made, include random effect of stimulus.<sup>23</sup>

3364    Don’t include lots of other stuff in your default model. You are doing  
3365    a randomized experiment, and the strength of randomized experiments  
3366    is that you don’t have to worry about confounding based on the popu-  
3367    lation (see chapter 1). So don’t put a lot of covariates in your default

Barr et al. 2013

3368 model—usually don’t put in any!<sup>24</sup>

3369 This default saturated model then represents a simple summary of your  
3370 experimental results. Its coefficients can be interpreted as estimates of  
3371 the effects of interest, and it can be used as the basis for inferences about  
3372 the relation of the experimental effect to the population using either  
3373 frequentist or Bayesian tools.

3374 Here’s a bit more guidance about this modeling strategy.

3375 1. **Preregister your model.** If you change your analysis approach af-  
3376 ter you see your data, you risk *p*-hacking—choosing an analysis  
3377 that biases the estimate of your effect of interest. As we discussed  
3378 in chapter 3 and as we will discuss in more detail in chapter 11,  
3379 one important strategy for minimizing this problem is to prereg-  
3380 ister your analysis.<sup>25</sup>

3381 2. **Visualize the model predictions against the observed data.** As  
3382 we’ll discuss in chapter 15, the “default model” for an experi-  
3383 ment should go alongside a “default visualization” known as the  
3384 **design plot** that similarly reflects the full design structure of the  
3385 experiment and any primary clusters. One way to check whether  
3386 a model fits your data is then to plot it on top of those data. Some-  
3387 times this combination of model and data can be as simple as a

<sup>24</sup> One corollary to having this kind of default perspective on data analysis: When you see an analysis that deviates substantially from the default, these deviations should provoke some questions. If someone drops a manipulation from their analysis, adds a covariate or two, or fails to control for some clustering in the data, did they deviate because of different norms in their subfield, or was there some other rationale? This line of reasoning sometimes leads to questions about the extent to which particular analytic decisions are post hoc and driven by the data (in other words, *p*-hacked). For an example, see the case study in chapter 11.

<sup>25</sup> A side benefit of preregistration is it makes you think through whether your experimental design is appropriate—that is, is there actually an analysis capable of estimating the effect you want from the data you intend to collect?

3388 scatter plot with a regression line. But seeing the model plotted  
3389 alongside the data can often reveal a mismatch between the two.  
3390 A model that does not describe the data very well is not a good  
3391 source of generalizable inferences!

3392 **3. Interpret the model predictions.** Once you have a model, don't  
3393 just read off the *p*-values for your coefficients of interest. Walk  
3394 through each coefficient, considering how it relates to your out-  
3395 come variable. For a simple two-group design like we've been  
3396 considering, the condition coefficient is the estimate of the causal  
3397 effect that you intended to measure! Consider its sign, its magni-  
3398 tude, and its precision (standard error or confidence interval).

3399 That said, there are some contexts in which it does make sense to de-  
3400 part from the default saturated model. For example, there may be in-  
3401 sufficient statistical power to estimate multiple interaction terms, or co-  
3402 variates might be included in the model to help handle certain forms of  
3403 missing data. The default model simply represents a very good starting  
3404 point.



Figure 7.3  
Example stimulus materials analogous  
to those used in Stiller, Goodman, and  
Frank (2015).

3405 7.4.2 *A worked example*

3406 All this advice may seem abstract, so let's put it into practice on a simple  
3407 example. For a change, let's look at an experiment that's not about tea  
3408 tasting. Here we'll consider data from an experiment testing preschool  
3409 children's language comprehension (Stiller, Goodman, and Frank 2015).

3410 In this experiment, two- to five-year-old children saw displays like the  
3411 one in figure 7.3. In the experimental condition, a puppet might say, for  
3412 example, "My friend has glasses! Which one is my friend?" The goal  
3413 was to measure how many children made the "pragmatic inference"  
3414 that the puppet's friend was the face with glasses and *no* hat.

3415 To estimate the effect, participants were randomly assigned to either the  
3416 experimental condition or to a control condition in which the puppet  
3417 had eaten too much peanut butter and couldn't talk, but they still had  
3418 to guess which face was his friend. There were also three other types  
3419 of experimental stimuli (houses, beds, and plates of pasta). Data from  
3420 this experiment consisted of 588 total observations from 147 children,  
3421 with all four stimuli presented to each child. The primary hypothesis of  
3422 this experiment was that that preschool children could make pragmatic  
3423 inferences by correctly inferring which of the three faces (for example)  
3424 the puppet was describing.

## CODE

If you want to follow along with this example, you'll have to load the example data and do a little bit of preprocessing (also covered in appendix D):

```
repo <- "https://raw.githubusercontent.com/langcog/experimentology/main"
sgf <- read_csv(file.path(repo, "data/tidyverse/stiller_scales_data.csv")) |>
  mutate(age_group = cut(age, 2:5, include.lowest = TRUE),
         condition = condition |>
  fct_recode("Experimental" = "Label", "Control" = "No Label"))
```

3425

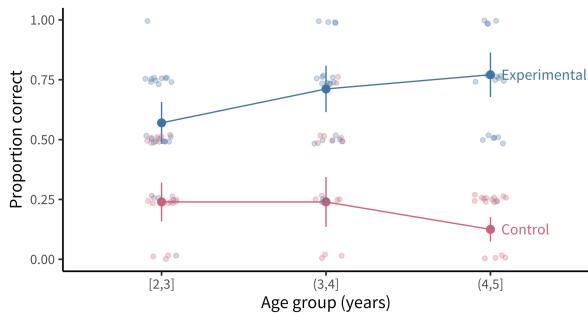


Figure 7.4

Data for Stiller, Goodman, and Frank (2015). Each point shows a single participant's proportion correct trials (out of four experimental stimuli) plotted by age group, jittered slightly to avoid overplotting. Larger points and associated confidence intervals show mean and 95% confidence intervals for each condition.

3426 This experimental design looks a lot like some versions of our tea-  
 3427 tasting experiment. We have one primary condition manipulation  
 3428 (the puppet provides information versus does not), presented between  
 3429 participants so that some participants are in the experimental condition  
 3430 and others are in the control condition. Our measurements are  
 3431 repeated within participants across different experimental stimuli.  
 3432 Finally, we have one important, preplanned covariate: children's age.  
 3433 The experimental data are plotted in figure 7.4.<sup>26</sup>

<sup>26</sup> Our sampling plan for this experiment was actually **stratified** across age, meaning that we intentionally recruited the same number of participants for each one-year age group—because we anticipated that age was highly correlated with children's ability to succeed in this task. We'll describe this kind of sampling in more detail in chapter 10.

3434 How should we go about making our default model for this dataset?<sup>27</sup>

3435 We simply include each of these design factors in a mixed effects model;

3436 we use a logistic link function for our mixed effects model (a **general-**

3437 **ized linear mixed effects model**) because we would like to predict cor-

3438 rect performance on each trial, which is a binary variable. So that gives

3439 us an effect of condition and age as a covariate. We further add an in-

3440 teraction between condition and age in case the condition effect varies

3441 meaningfully across groups. Finally, we add random effects of partici-

3442 pant,  $\gamma_i$ , and experimental item,  $\gamma_t$ .<sup>28</sup>

3443 The resulting model looks like this:

$$\text{logit}(E[Y_{it}]) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}X_{i2} + \gamma_i + \delta_t$$

3444 Let's break this complex equation down from left to right:

- 3445 –  $\text{logit}(E[Y_{it}])$  says that we are predicting a logistic function of
- 3446  $E[Y_{it}]$  (where  $Y_{it}$  indicates whether child  $i$  was correct on trial
- 3447  $t$ ).
- 3448 –  $\beta_0$  is the **intercept**, our estimate of the average log-odds (i.e., the
- 3449 log of the odds ratio) of correct responses for participants in the
- 3450 control condition.
- 3451 –  $\beta_1 X_{i1}$  is the condition predictor.  $\beta_1$  represents the change in log-
- 3452 odds associated with being in the experimental condition (the

<sup>27</sup> This experiment was not preregistered, but the paper includes a separate replication dataset with the same analysis.

<sup>28</sup> As discussed above, this is a tricky decision point; we could very reasonably have added random slopes as well.

causal effect of interest!), and  $X_{i1}$  is an indicator variable that is 1 if child  $i$  is in the experimental condition and 0 for the control condition. Multiplying  $\beta_1$  by this indicator means that the predictor has the value 0 for participants in the control condition and  $\beta_1$  for those in the experimental condition.

- $\beta_2 X_{i2}$  is the age predictor.  $\beta_2$  represents the difference in log odds associated with one additional year of age for participants in the control condition<sup>29</sup> and  $X_{i2}$  is the age for each participant.<sup>30</sup>
- $\beta_3 X_{i1} * X_{i2}$  is the interaction between experimental condition and age.  $\beta_3$  represents the difference in log odds (i.e., the log of the odds ratio) that is associated with being one year older *and* in the experimental condition versus the control condition. This term is multiplied by both each child's age *and* the condition indicator  $X_i$ .
- $\gamma_i$  is the random intercept for participant  $i$ , capturing individual variation in the odds of success across trials.
- $\gamma_t$  is the random intercept for stimulus  $t$ , capturing variation in the odds of success across the four different stimuli.

Table 7.2

Estimated effects for our generalized linear mixed effects model on data from Stiller, Goodman, and Frank (2015).

term	estimate	conf.int	statistic	p.value
Control condition	0.80	[0.42, 1.18]	4.16	< .001
Age (years)	0.55	[0.21, 0.88]	3.19	.001

<sup>29</sup> The age coefficient is a **simple effect**, meaning it is the effect of age in the control condition only. That's because we have dummy coded the condition predictor. If we wanted the average age effect (the **main effect**) then we would need to use contrast coding, per the note in the “Interactions” section above.

<sup>30</sup> We have **centered** our age predictor in this example so that all estimates from our model are for the average age of our participants. Centering is a good practice for modeling continuous predictors because it increases the interpretability of other parts of the model. For example, because age is centered in this model, the intercept  $\beta_0$  can be interpreted as the predicted odds of a correct trial for a participant in the control condition at the average age.

Table 7.2

Estimated effects for our generalized linear mixed effects model on data from Stiller, Goodman, and Frank (2015).

term	estimate	conf.int	statistic	p.value
Expt condition	-2.26	[-2.70, -1.82]	-10.07	< .001
Age (years) * Expt condition	-0.92	[-1.43, -0.42]	-3.60	< .001

## CODE

To fit the model described above, the first step is to prepare your predictors. In this case, we center the age predictor.

```
sgf$age_centered <- scale(sgf$age, center = TRUE, scale = FALSE)
```

Again we use the `lme4` package, this time with the `glmer()` function.

Again we have to specify our link function, just like in a standard GLM, by choosing the distribution family.

```
mod <- glmer(correct ~ age_centered * condition + (1|subid) + (1|item),
               family = "binomial", data = sgf)
```

You can see a summary of the fitted model using `summary(mod)` as before.

The only big difference from `lm()` is that here you can extract both fixed and random effects (with `fixef(mod)` and `ranef(mod)`, respectively).

3471

Let's estimate this model and see how it looks. We'll focus here on interpretation of the so-called **fixed effects** (the main predictors), as opposed to the participant and item random effects.<sup>31</sup> Table 7.2 shows the coefficients. Again, let's walk through each:

<sup>31</sup> Participant means are estimated to have a standard deviation of 0.23 (in log-odds), while items have a standard deviation of 0.27. These indicate that both of our random effects capture meaningful variation.

3471

- 3476 – The **intercept** (control condition estimate) is  $\hat{\beta} = 0.80$ , 95% CI  
3477 [0.42, 1.18],  $z = 4.16$ ,  $p < 0.001$ . This estimate reflects that the  
3478 log-odds of a correct response for an average-age participant in  
3479 the control condition is 0.8, which corresponds to a probability  
3480 of 0.69. If we look at figure 7.4, that estimate makes sense: 0.69  
3481 seems close to the average for the control condition.
- 3482 – The **age effect** estimate is  $\hat{\beta} = 0.55$ , 95% CI [0.21, 0.88],  $z = 3.19$ ,  
3483  $p = 0.001$ . This means there is a slight decrease in the log-odds  
3484 of a correct response for older children in the control condition.  
3485 Again, looking at figure 7.4, this estimate is interpretable: we see a  
3486 small decline in the probability of a correct response for the oldest  
3487 age group.
- 3488 – The key experimental condition estimate then is  $\hat{\beta} = -2.26$ , 95%  
3489 CI [-2.70, -1.82],  $z = -10.07$ ,  $p < 0.001$ . This estimate means  
3490 that the log-odds of a correct response for an average-age partic-  
3491 ipant in the experimental condition is the sum of the estimates  
3492 for the control (intercept) and the experimental conditions: 0.8  
3493 + -2.26, which corresponds to a probability of 0.19. Grounding  
3494 our interpretation in figure 7.4, this estimate corresponds to the  
3495 average value for the experimental condition.
- 3496 – Finally, the **interaction** of age and condition is  $\hat{\beta} = -0.92$ , 95% CI

3497  $[-1.43, -0.42]$ ,  $z = -3.60$ ,  $p < 0.001$ . This positive coefficient  
3498 reflects that with every year of age, the difference between control  
3499 and experimental conditions grows.

3500 In sum, this model suggests that there was a substantial difference in  
3501 performance between experimental and control conditions, in turn sup-  
3502 porting the hypothesis that children in the sampled age group can per-  
3503 form pragmatic inferences above chance.

3504 This example illustrates the “default saturated model” framework that  
3505 we recommend—the idea that a single regression model corresponding  
3506 to the design of the experiment can yield an interpretable estimate of  
3507 the causal effect of interest, even in the presence of other sources of  
3508 variation.

#### 🔍 DEPTH

*When does it make sense to include covariates in a  
model?*

Let's come back to one piece of advice that we gave above about making a “default” model of an experiment: not including covariates. This advice can seem surprising. Many demographic factors are of interest to psychologists and other behavioral scientists, and in observational studies these factors will almost always be related to important life outcomes.

So, why not put them into our experimental models? After all, we did include age in our worked example above!

Well, if you have one or at most a small handful of covariates that you believe are meaningfully related to the outcome, you *can* plan in advance to put them in your model. If you think that your effect is likely to be moderated by a specific demographic characteristic—as we did with age in our developmental example above—then this inclusion can be quite useful.

Further, including covariates can increase the precision of your estimates by reducing “noise” in your outcome, if you hypothesize that they interact. What’s surprising though is how *little* this adjustment does to increase your overall precision unless the correlation between covariate and outcome is very strong.

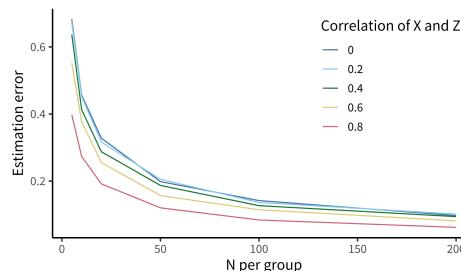


Figure 7.5  
Decreases in estimation error due to adjusting for covariates, plotted by the N participants in each group and the correlation between the outcome (X) and the covariate (Z).

Figure 7.5 shows the results of a simple simulation investigating the relationship between estimation error and the inclusion of covariates. Only when the correlation between covariate and outcome (e.g., age and tea rating) is greater than  $r = 0.6$  to  $r = 0.8$  does this adjustment really help.

That said, there are quite a few reasons not to include covariates. These motivate our recommendation to skip them in your default model unless you have very strong theory-based expectations for either (A) a correlation with the outcome or (B) a strong moderation relationship.

The first reason not to include covariates is simply because we don't need to. Because randomization cuts causal links, our experimental estimate is an unbiased estimate of the causal effect of interest (at least for large samples). We are guaranteed that, in the limit of many different experiments, even though people with different ages will be in the different tea-tasting conditions, this source of variation will be averaged out. Actually, including unnecessary covariates into models (slightly) decreases the probability that the model can detect a true effect (that is, it decreases statistical precision and power). Just by chance, covariates can "soak up" variation in the outcome, leaving less to be accounted for by the true effect!

The second reason is that you can actually compromise your causal inference by including some covariates, particularly those that are collected *after* randomization. The logic of randomization is that you cut all causal links between features of the sample and the condition manipulation. But you can "uncut" these links by accident by adding variables into your model that are related to group status. This problem is generically called **conditioning on post-treatment variables**, and a full discussion of is out of the scope of this book, but it's something to avoid (and read up on if you're worried about it, see [Montgomery, Nyhan, and Torres 2018](#)).

Finally, one of the standard justifications for adding covariates—because your groups are unbalanced—is actually ill-founded as well. People often talk about “unhappy randomization”: you randomize to the different tea-tasting groups, for example, but then it turns out the mean age is a bit different between groups. Then you do a *t*-test or some other statistical test and find out that you actually have a significant age difference. This practice makes no sense! Because you randomized, you know that the difference in ages occurred by chance. Further, incidental demographic differences between groups are unlikely to be important unless that characteristic is highly correlated with the outcome (see above). Instead, if the sample size is small enough that meaningfully large incidental differences could arise in important confounders, then it is preferable to **stratify** on that confounder at the outset—we’ll have lot more to say about this issue in chapter 10.

So these are our options: if a covariate is known to be very strongly related to our outcome, we can include it in our default model. Otherwise, we avoid a lot of trouble by leaving covariates out.

3512

#### 3513 7.4.1 Robustness checks and the multiverse

3514 Using the NHST statistical testing approach that has been common in

3515 the psychology literature, even a simple two-factor experimental design

3516 affords a host of different *t*-tests and ANOVAs,<sup>32</sup> offering many oppor-

3517 tunities for *p*-hacking and selective reporting. We’ve been advocating

here instead for a “default model” approach in which you preplan and preregister a single regression model that captures the planned features of your experimental design, including manipulations and sources of clustering. This approach can help you to navigate some of the complexity of data analysis by having a standard approach that you take in almost every case.

Not every dataset will be amenable to this approach, however. For complex experimental designs or unusual measures, sometimes it can be hard to figure out how to specify or fit the default saturated model. And especially, in these cases, the choice of model can make a big difference to the magnitude of the reported effect. To quantify variability in effect size due to model choice, “Many Analysts” projects have asked a set of teams to approach a dataset using different analysis methods. The result from these projects has been that there is substantial variability in outcomes depending on what approach is taken (Silberzahn et al. 2018; Botvinik-Nezer et al. 2020).<sup>33</sup>

**Robustness analysis** (also sometimes called “sensitivity analysis” or “multiverse analysis,” which sounds cooler) is a technique for addressing the possibility that an individual analysis overestimates or underestimates a particular effect by chance (Steegen et al. 2016). The general idea is that analysts explore a space of different possible analyses.

<sup>33</sup> To be fair, often the analytic questions being investigated in “Many Analysts” projects are more complex than the simple experiments we recommend doing, and there is debate about how much true variability these investigations reveal (Breznau et al. 2022; Mathur, Covington, and VanderWeele 2023).

3539 In its simplest form, alternative model specifications can be reported in  
3540 a supplement; more sophisticated versions of the idea call for averaging  
3541 estimates across a range of possible specifications and reporting this  
3542 average as the primary effect estimate.

3543 The details of this kind of analysis will vary depending on what you  
3544 are worried about your model being sensitive to. One analyst might  
3545 be concerned about the effects of adding different covariates; another  
3546 might be using a Bayesian framework and be concerned about sensitivity  
3547 to particular prior values. If you get similar results across many different  
3548 specifications, you can sleep better at night. The primary principle to  
3549 take home is a bit of humility about our models. Any given model is  
3550 likely wrong in some of its details. Investigating the sensitivity of your  
3551 estimates to the details of your model specification is a good idea.

## 3552 *7.5 Chapter summary: Models*

3553 In the last three chapters, we have spelled out a framework for data  
3554 analysis that focuses on our key experimental goal: a measurement of a  
3555 particular causal effect. We began with basic techniques for estimating  
3556 effects and making inferences about how these effects estimated from a  
3557 sample can be generalized to a population. This chapter showed how  
3558 these ideas naturally give rise to the idea of making models of data,

3559 which allow estimation of effects in more complex designs. Simple re-  
3560 gression models, which are formally identical to other inference meth-  
3561 ods in the most basic case, can be extended with the generalized linear  
3562 model as well as with mixed effects models. Finally, we ended with  
3563 some guidance on how to build a “default model”—an (often preregis-  
3564 tered) regression model that maps onto your experimental design and  
3565 provides the primary estimate of your key causal effect.



### DISCUSSION QUESTIONS

1. Choose a paper that you have read for your research and take a look at the statistical analysis. Does the reporting focus more on hypothesis testing or on estimating effect sizes?
2. We focused here on the linear model as a tool for building models, contrasting this perspective with the common “statistical testing” mindset. But—here’s the mind-blowing thing—most of those statistical tests are special cases of the linear model anyway. Take a look at this extended meditation on the equivalences between tests and models: [https://lindeloev.github.io/tests-as-linear/#9\\_teaching\\_materials\\_and\\_a\\_course\\_outline](https://lindeloev.github.io/tests-as-linear/#9_teaching_materials_and_a_course_outline). If the paper you chose for question 1 used tests, could their tests be easily translated to models? How would the use of a model-based perspective change the results section of the paper?
3. Take a look at this cool visualization of hierarchical (mixed effect)

models: <http://mfviz.com/hierarchical-models>. In your own research, what are the most common units that group together your observations?

3567

## READINGS

- An opinionated practical guide to regression modeling and data description: Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020). *Regression and Other Stories*. Cambridge University Press. Free online at <https://avehtari.github.io/ROS-Examples>.
- A more in-depth introduction to the process of developing Bayesian models of data that allow for estimation and inference in complex datasets: McElreath, Richard (2018). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC. Free materials available at <https://xcelab.net/rm/statistical-rethinking>.

3568

III

3569

## PLANNING

3570

<sup>3571</sup> *References*

- Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. “Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal.” *Journal of Memory and Language* 68 (3): 255–278.
- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software* 67 (1): 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Bie, Ruofan, Sébastien Haneuse, Nathan Huey, Jonathan Schildcrout, and Glen McGee. 2021. “Fitting Marginal Models in Small Samples: A Simulation Study of Marginalized Multilevel Models and Generalized Estimating Equations.” *Statistics in Medicine* 40 (24): 5298–5312.
- Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. “Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams.” *Nature* 582 (7810): 84–88.
- Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, et al. 2022. “Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty.” *Proceedings of the National Academy of Sciences* 119 (44): e2203150119.
- Bürkner, Paul-Christian. 2021. “Bayesian Item Response Modeling in R with brms and Stan.” *Journal of Statistical Software* 100 (5): 1–54. <https://doi.org/10.18637/jss.v100.i05>.
- Davis, Matthew J. 2010. “Contrast Coding in Multiple Regression Analysis: Strengths, Weaknesses, and Utility of Popular Coding Structures.” *Journal*

- of Data Science* 8 (1): 61–73.
- Galton, Francis. 1877. *Typical Laws of Heredity*. Royal Institution of Great Britain.
- Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press.
- Mancl, Lloyd A., and Timothy A. DeRouen. 2001. “A Covariance Estimator for GEE with Improved Small-Sample Properties.” *Biometrics* 57 (1): 126–134.
- Mathur, Maya B., Christian Covington, and Tyler J. VanderWeele. 2023. “Variation across Analysts in Statistical Significance, Yet Consistently Small Effect Sizes.” *Proceedings of the National Academy of Sciences* 120 (3): e2218957120.
- McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.
- Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. “How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It.” *American Journal of Political Science* 62 (3): 760–775.
- Robinson, David, Alex Hayes, and Simon Couch. 2023. *broom: Convert Statistical Objects into Tidy Tibbles*. <https://CRAN.R-project.org/package=broom>.
- Silberzahn, Raphael, Eric L. Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, et al. 2018. “Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results.” *Advances in Methods and Practices in Psychological Science* 1 (3): 337–356.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel.

2016. "Increasing Transparency through a Multiverse Analysis." *Perspectives on Psychological Science* 11 (5): 702–712. <https://doi.org/10.1177/1745691616658637>.

Stiller, Alex J., Noah D. Goodman, and Michael C. Frank. 2015. "Ad-Hoc

Implicature in Preschool Children." *Language Learning and Development* 11 (2): 176–190.

## 8 MEASUREMENT

3575



### LEARNING GOALS

- Discuss the reliability and validity of psychological measures
- Reason about trade-offs between different measures and measure types
- Identify the characteristics of well-constructed survey questions
- Articulate risks of measurement flexibility and the costs and benefits of multiple measures

3576

3577 In the previous section of the book, we described a set of measurement-focused statistical techniques for quantifying (and maximizing) our precision. In this next set of three chapters focusing on planning experiments, we will develop our toolkit for designing the measures (this chapter), design manipulations (chapter 9), and sampling (chapter 10) strategies that will allow us to create and evaluate experiments. These chapters form a core part of our approach to “experimentology”: a set

3578

3579

3580

3581

3582

3583

3584 of decisions to REDUCE BIAS, maximize MEASUREMENT PRECISION, and  
3585 assess GENERALIZABILITY. Let's begin with measurement.

3586 Throughout the history of science, advances in measurement have gone  
3587 hand in hand with advances in knowledge.<sup>1</sup> Telescopes revolutionized  
3588 astronomy, microscopes revolutionized biology, and patch clamping  
3589 revolutionized physiology. But measurement isn't easy. Even the  
3590 humble thermometer, allowing reliable measurement of temperature,  
3591 required centuries of painstaking effort to perfect (Chang 2004). Psy-  
3592 chology and the behavioral sciences are no different—we need reliable  
3593 instruments to measure the things we care about. In this next section  
3594 of the book, we're going to discuss the challenges of measurement in  
3595 psychology, and the properties that distinguish good instruments from  
3596 bad.

3597 What does it mean to measure something? Intuitively, we know that a  
3598 ruler measures the quantity of length and a scale measures the quantity  
3599 of mass (Kisch 1965). As we discussed in chapter 2, those quantities are  
3600 latent (unobserved). Individual measurements, in contrast, are manifest:  
3601 they are observable to us. What does it mean to measure a psychological  
3602 construct—a hypothesized theoretical quantity inside the head?

3603 We first have to keep in mind that not every measure is equally precise.  
3604 This point is obvious when you think about physical measurement in-

<sup>1</sup> As such, measurement is a perennially controversial topic in philosophy of science. For an overview of competing frameworks, see Tal (2020) or Maul, Irribarra, and Wilson (2016), which focuses specifically on measurement in psychology.

3605 struments: a caliper will give you a much more precise estimate of thick-  
3606 ness than a ruler will. One way to see that the measurement is more  
3607 precise is by repeating it a bunch of times. The measurements from the  
3608 caliper will likely be more similar to one another, reflecting the fact that  
3609 the amount of error in each individual measurement is smaller. We can  
3610 do the same thing with a psychological measurement—repeat and assess  
3611 variation—though as we’ll see below it’s a little trickier. Measurement  
3612 instruments that have less error are called more **reliable** instruments.<sup>2</sup>

3613 Second, psychological measurements do not directly reflect latent the-  
3614 oretical constructs of interest, quantities like happiness, intelligence, or  
3615 language processing ability. And unlike quantities like length and mass,  
3616 there is often disagreement in psychology about what the right theoret-  
3617 ical quantities are. Thus, we have to measure an observable behavior—  
3618 our operationalization of the construct—and then make an argument  
3619 about how the measure relates to a proposed construct of interest (and  
3620 sometimes whether the construct really exists at all). This argument is  
3621 about the **validity** of our measurements.<sup>3</sup>

3622 These two concepts, reliability and validity, provide a conceptual toolkit  
3623 for assessing a psychological measurement and how well it serves the  
3624 researcher’s goal.

<sup>2</sup> Is reliability the same as **precision**? Yes, more or less. Confusingly, different fields call these concepts different things (there’s a helpful table of these names in [Brandmaier et al. 2018](#)). Here we’ll talk about reliability as a property of instruments specifically while using the term precision to talk about the measurements themselves.

<sup>3</sup> We are also going to talk in chapter 9 about the validity of manipulations. The way you identify a causal effect on some measure is by operationalizing some construct as well. To identify causal effects, we must link a particular construct of interest to something we can concretely manipulate in an experiment, like the stimuli or instructions.

3625 **8.1 Reliability**

3626 Reliability is a way of describing the extent to which a measure yields  
 3627 signal relative to noise. Intuitively, if there's less noise, then there will be  
 3628 more similarity between different measurements of the same quantity,  
 3629 illustrated in figure 8.1 as a tighter grouping of points on the bulls-eye.

3630 But how do we measure signal and noise?

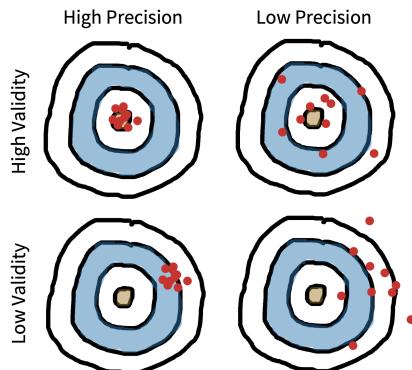


Figure 8.1  
 Reliability and validity visualized. The reliability of an instrument is its expected precision. The bias of measurements from an instrument also provide a metaphor for its validity.

 CASE STUDY

### *A reliable and valid measure of children's vocabulary*

Anyone who has worked with little children, or had children of their own, can attest to how variable their early language is. Some children speak clearly and produce long sentences from an early age, while others struggle; this variation appears to be linked to later school outcomes (Marchman and Fernald 2008). Thus, there are many reasons why you'd want to make precise measurements of children's early language ability as a latent construct of interest.

Because bringing children into a lab can be expensive, one popular alter-

native option for measuring child language is the MacArthur Bates Communicative Development Inventory (CDI for short), a form that asks parents to mark words that their child says or understands. CDI forms are basically long checklists of words. But is parent report a reliable or valid measure of children's early language?

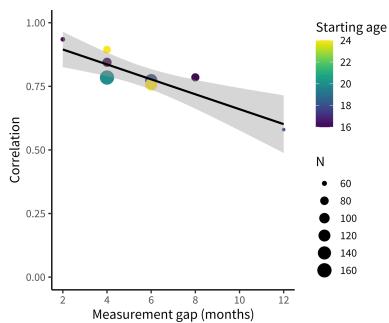


Figure 8.2  
Longitudinal (test-retest) correlations between a child's score on one administration of the CDI and another one several months later. Based on Frank et al. (2021).

As we'll see below, one way to measure the reliability of the CDI is to compute the correlation between two different administrations of the form for the same child. Unfortunately, this analysis has one issue: the longer you wait between observations the more the child has changed! Figure 8.2 displays these correlations for two CDIs, showing how correlations start off high and drop off as the gap between observations increases (Frank et al. 2021).

Given that CDI forms are relatively reliable instruments, are they valid? That is, do they really measure the construct of interest, namely children's early language ability? Bornstein and Haynes (1998) collected many different measures of children's language—including the Early Language In-

ventory (ELI; an early CDI form) and other “gold standard” measures like transcribed samples of children’s speech. CDI scores were highly correlated with all the different measures, suggesting that the CDI was a valid measure of the construct.

The combination of reliability and validity evidence suggests that CDIs are a useful (and relatively inexpensive) source of data about children’s early language, and indeed they have become one of the most common assessments for this age group!

## 3634 8.1.1 Measurement scales

3635 In the physical sciences, it's common to measure the precision of an  
 3636 instrument using its coefficient of variation (Brandmaier et al. 2018):

$$CV = \frac{\sigma_w}{\mu_w}$$

3637 where  $\sigma_w$  is the standard deviation of the measurements within an in-  
 3638 dividual and  $\mu_w$  is the mean of those measurements (figure 8.3).

3639 Imagine we measure the height of a person five times, resulting in mea-  
 3640 surements of 171 cm, 172cm, 171 cm, 173 cm, and 172 cm. These are  
 3641 the combination of the person's true height (we assume they have one!)  
 3642 and some **measurement error**. Now we can use these measurements  
 3643 to compute the coefficient of variation, which is 0.005, suggesting very  
 3644 limited variability relative to the overall quantity being measured. Why  
 3645 can't we just do this same thing with psychological measurements?

3646 Thinking about this question takes us on a detour through the differ-  
 3647 ent kinds of measurement scales used in psychological research (Stevens  
 3648 1946). The height measurements in our example are on what is known  
 3649 as a **ratio scale**: a scale in which numerical measurements are equally  
 3650 spaced and on which there is a true zero point. These scales are com-  
 3651 mon for physical quantities but somewhat less frequent in psychology  
 3652 (with reaction times as a notable exception). More common are **interval**

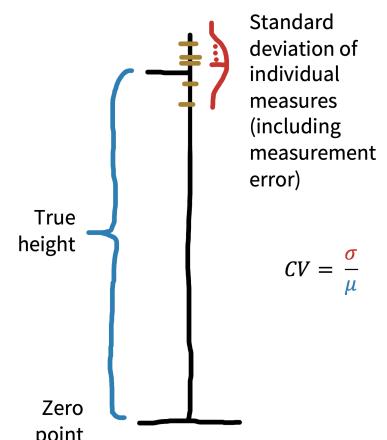
**Measuring reliability with a true zero**

Figure 8.3  
 Computing the coefficient of variation (CV).

3653 scales, in which there is no true zero point. For example, IQ (and other  
 3654 standardized scores) are intended to capture interval variation on some  
 3655 dimension but zero is meaningless—an IQ of zero does not correspond  
 3656 to any particular interpretation.<sup>4</sup>

3657 **Ordinal** scales are also often used. These are scales that are ordered but  
 3658 are not necessarily spaced equally. For example, levels of educational  
 3659 achievement (“elementary,” “high school,” “some college,” “college,”  
 3660 “graduate school”) are ordered, but there is no sense in which “high  
 3661 school” is as far from “elementary” as “graduate school” is from “col-  
 3662 lege.” The last type in Stevens’s hierarchy is **nominal** scales, in which  
 3663 no ordering is possible either. For example, race is an unordered scale in  
 3664 which multiple categories are present but there is no inherent ordering  
 3665 of these categories. The hierarchy is shown in table 8.1.

Table 8.1  
 Scale types and their associated operations and statistics (Stevens 1946).

Scale	Definition	Operations	Statistics
Nominal	Unordered list	Equality	Mode
Ordinal	Ordered list	Greater than or less than	Median
Interval	Numerical	Equality of intervals	Mean, SD
Ratio	Numerical & zero	Equality of ratios	Coefficient of variation

3666 Critically, different summary measures work for each scale type. If you  
 3667 have an unordered list like a list of options for a question about race on

<sup>4</sup> It can actually be shown in a suitably rigorous sense that ratio and interval scales (and another lying in between) are the *only* scales possible for the real numbers (Narens and Luce 1986).

3668 a survey, you can present the modal response (the most likely one). It  
3669 doesn't even make sense to think about what the median was—there's  
3670 no ordering! For ordered levels of education, a median is possible but  
3671 you can't compute a mean. And for interval variables like “number of  
3672 correct answers on a math test,” you can compute a mean and a standard  
3673 deviation.<sup>5</sup>

3674 Now we're ready to answer our initial question about why we can't  
3675 quantify reliability using the coefficient of variation. Unless you have a  
3676 ratio scale with a true zero, you can't compute a coefficient of variation.  
3677 Think about it for IQ scores: currently, by convention, standardized IQ  
3678 scores are set to have a mean of 100. If we tested someone multiple times  
3679 and found the standard deviation of their test scores was four points,  
3680 then we could estimate the precision of their measurements as “CV” of  
3681  $4/100 = 0.04$ . But since IQ of zero isn't meaningful, we could just set  
3682 the mean IQ for the population to 200. Our test would be the same,  
3683 and so the CV would be  $4/200 = 0.02$ . On that logic, we just doubled  
3684 the precision of our measurements by rescaling the test! That doesn't  
3685 make any sense.

<sup>5</sup> You might be tempted to think that “number of correct answers” is a ratio variable—but is zero really meaningful? Does it truly correspond to “no math knowledge” or is it just a stand-in for “less math knowledge than this test requires”?

 DEPTH

### *Early controversies over psychological measurement*

Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of ... measurement.

—Cattell (1890, p. 373)

It is no coincidence that the founders of experimental psychology were obsessed with measurement (Heidelberger 2004). It was viewed as the primary obstacle facing psychology on its road to becoming a legitimate quantitative science. For example, one of the final pieces written by Hermann von Helmholtz (Wilhelm Wundt's doctoral advisor) was a 1887 philosophical treatise titled “Zahlen und Messen” (“Counting and Measuring,” see Darrigol 2003). In the same year, Fechner (1987 [1887]) explicitly grappled with the foundations of measurement in “Über die psychischen Massprincipien” (“On psychic measurement principles”).

Many of the early debates over measurement revolved around the emerging area of *psychophysics*, the problem of relating objective, physical stimuli (e.g., light, sound, pressure) to the subjective sensations they produce in the mind. For example, Fechner (1860) was interested in a quantity called the “just noticeable difference”—the smallest change in a stimulus that can be discriminated by our senses. He argued for a lawful (logarithmic) relationship: a logarithmic change in the intensity of, say, brightness corresponded to a linear change in the intensity people reported (up to

some constant). In other words, sensation was *measurable* via instruments like just noticeable difference.

It may be surprising to modern ears that the basic claim of measurability was controversial, even if the precise form of the psychophysical function would continue to be debated. But this claim led to a deeply rancorous debate, culminating with the so-called Ferguson Committee, formed by the British Association for the Advancement of Science in 1932 to investigate whether such psychophysical procedures could count as quantitative “measurements” of anything at all (Moscati 2018). It was unable to reach a conclusion, with physicists and psychologists deadlocked:

Having found that individual sensations have an order, they [some psychologists] assume that they are *measurable*. Having travestied physical measurement in order to justify that assumption, they assume that their sensation intensities will be related to stimuli by numerical laws ... which, if they mean anything, are certainly false.

—Ferguson et al. (1940, p. 347)

The heart of the disagreement was rooted in the classical definition of quantity requiring a strictly *additive* structure. An attribute was only considered measurable in light of a meaningful concatenation operation. For example, weight was a measurable attribute because putting a bag of three rocks on a scale yields the same number as putting each of the three rocks on separate scales and then summing up those numbers (in philosophy

of science, attributes with this concatenation property are known as “extensive” attributes, as opposed to “intensive” ones). Norman Campbell, one of the most prominent members of the Ferguson Committee, had recently defined *fundamental measurement* in this way (e.g., [Campbell 1928](#)), contrasting it with *derived measurement*, which involved computing some function based on one or more fundamental measures. According to the physicists on the Ferguson Committee, measuring mental sensations was impossible because they could never be grounded in any *fundamental scale* with this kind of additive operation. It just didn’t make sense to break up people’s sensations into parts the way we would weights or lengths: they didn’t come in “amounts” or “quantities” that could be combined ([Catell 1962](#)). Even the intuitive additive logic of Donders’s ([1969 \[1868\]](#)) “method of subtraction” for measuring the speed of mental processes was viewed skeptically on the same grounds by the time of the committee (e.g., in an early textbook, [Woodworth \(1938, p. 83\)](#) claimed, “We cannot break up the reaction into successive acts and obtain the time for each act”).

The primary target of the Ferguson Committee’s investigation was the psychologist S. S. Stevens, who had claimed to measure the sensation of loudness using psychophysical instruments. Exiled from classical frameworks of measurement, he went about developing an alternative “operational” framework ([Stevens 1946](#)), where the classical ratio scale recognized by physicists was only one of several ways of assigning numbers to things (see table 8.1 above). Stevens’s framework quickly spread, leading

to an explosion of proposed measures. However, operationalism remains controversial outside psychology (Michell 1999). The most extreme version of Stevens's (1946, p. 677) stance ("Measurement ... is defined as the assignment of numerals to objects or events according to rules") permits researchers to *define* constructs operationally in terms of a measure (Hardcastle 1995). For example, one may say that the construct of intelligence is simply *whatever it is* that IQ measures. It is then left up to the researcher to decide which scale type their proposed measure should belong to.

In chapter 2, we outlined a somewhat different view, closer to a kind of constructive realism (Giere 2004; Putnam 2000). Psychological constructs like happiness are taken to exist independent of any given operationalization, putting us on firmer ground to debate the pros and cons associated with different ways of measuring the same construct. In other words, we are not free to assign numbers however we like. Whether a particular construct or quantity *is measurable* on a particular scale should be treated as an empirical question.

The next major breakthrough in measurement theory emerged with the birth of mathematical psychology in the 1960s, which aimed to put psychological measurement on more rigorous foundations. This effort culminated in the three-volume Foundations of Measurement series (Krantz et al. 1971; Suppes et al. 1989; Luce et al. 1990), which has become the canonical text for every psychology student seeking to understand measurement in the nonphysical sciences. One of the key breakthroughs was to shift the burden from measuring (additive) constructs themselves to

measuring (additive) *effects* of constructs in conjunction with one another:

When no natural concatenation operation exists, one should try to discover a way to measure factors and responses such that the ‘effects’ of different factors are additive.

—Luce and Tukey (1964, p. 4)

This modern viewpoint broadly informs the view we describe here.

3690

### 3691 8.1.1 Measuring reliability

3692 So then, how do we measure signal and noise when we don’t have a true  
3693 zero? We can still look at the variation between repeated measurement,  
3694 but rather than comparing that variation between measurements to the  
3695 mean, we can compare it to some other kind of variation—for exam-  
3696 ple, variation between people. In what follows, we’ll discuss reliability  
3697 on interval scales, but many of the same tools have been developed for  
3698 ordinal and nominal scales.

3699 Imagine that you are developing an instrument to measure some cogni-  
3700 tive ability. We assume that every participant has a true ability,  $t$ , just  
3701 the same way that they have a true height in the example above. Ev-  
3702 ery time we measure this true ability with our instrument, however, it  
3703 gets messed up by some measurement error. Let’s specify that error is

3704 normally distributed with a mean of zero—so it doesn't bias the mea-  
3705 surements, it just adds noise. The result is our observed score,  $o$ .<sup>6</sup>

3706 Taking this approach, we could define a relative version of the coeffi-  
3707 cient of variation. The idea is that the reliability of a measurement is  
3708 the amount of variance attributable to the true score variance (signal),  
3709 rather than the observed score variance (which includes noise). If  $\sigma_t^2$  is  
3710 the variance of the true scores and  $\sigma_o^2$  is the variance of the observed  
3711 scores, then this ratio is:

$$R = \frac{\sigma_t^2}{\sigma_o^2}.$$

3712 When noise is high, then the denominator is going to be big and  $R$  will  
3713 go down to 0; when noise is low, the numerator and the denominator  
3714 will be almost the same and  $R$  will approach 1.

3715 This all sounds great, except for one problem: we can't compute re-  
3716 liability using this formula without knowing the true scores and their  
3717 variance. But if we did, we wouldn't need to measure anything at all!

3718 There are two main approaches to computing reliability from data.  
3719 Each of them makes an assumption that lets you circumvent the  
3720 fundamental issue that we only have access to observed scores and not  
3721 true scores. Let's think these through in the context of a math test.

3722 **Test-retest reliability.** Imagine you have two parallel versions of your

<sup>6</sup> The approach we use to introduce this set of ideas is called **classical test theory**. There are other—more modern—alternative approaches, but CTT (as it's called) is a good starting point for thinking through the concepts.

3723 math test that are the same difficulty. Hence, you think a student's score  
 3724 on either one will reflect the same true score, modulo some noise. In  
 3725 that case, you can use these two sets of observed scores ( $o_1$  and  $o_2$ ) to  
 3726 compute the reliability of the instrument by simply computing the cor-  
 3727 relation between them ( $\rho_{o_1, o_2}$ ). The logic is that, if both variants reflect  
 3728 the same true score, then the shared variance (covariance in the sense of  
 3729 chapter 5) between them is just  $\sigma_t^2$ , the true score variance, which is the  
 3730 variable that we wanted but didn't have. Test-retest reliability is thus a  
 3731 very convenient way to measure reliability (figure 8.4).

3732 **Internal reliability.** If you don't have two parallel versions of the test, or  
 3733 you can't give the test twice for whatever reason, then you have another  
 3734 option. Assuming you have multiple questions on your math test (which  
 3735 is a good idea!), then you can split the test in pieces and treat the scores  
 3736 from each of these subparts as parallel versions. The simplest way to  
 3737 do this is to split the instrument in half and compute the correlation  
 3738 between participants' scores on the two halves—this quantity is called  
 3739 **split half reliability.**<sup>7</sup>

3740 Another method for computing the internal reliability (the **consistency**  
 3741 of a test) is to treat each test item as a subinstrument and compute the av-  
 3742 erage split-half correlation over all splits. This method yields the statis-  
 3743 tic **Cronbach's  $\alpha$**  ("alpha").  $\alpha$  is a widely reported statistic, but it is also

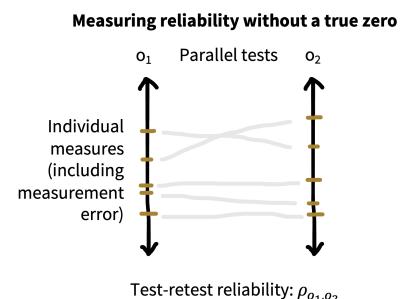


Figure 8.4  
 Computing test-retest reliability.

7 The problem is that each half is ...  
 half as long as the original instrument.  
 To get around this, there is a correction  
 called the Spearman-Brown correction  
 that can be applied to estimate the ex-  
 pected correlation for the full-length in-  
 strument. You also want to make sure  
 that the test doesn't get harder from the  
 beginning to the end. If it does, you may  
 want to use the even-numbered and odd-  
 numbered questions as the two parallel  
 versions.

3744 widely misinterpreted (Sijtsma 2009). First, it is actually a lower bound  
3745 on reliability rather than a good estimate of reliability itself. And second,  
3746 it is often misinterpreted as evidence that an instrument yields scores that  
3747 are “internally consistent,” which it does not; it’s not an accurate sum-  
3748 mary of dimensionality.  $\alpha$  is a standard statistic, but it should be used  
3749 with caution.

3750 One final note: these tools often get used for observers’ ratings of the  
3751 same stimulus (**inter-rater or inter-annotator reliability**), say for exam-  
3752 ple when you have two coders rate how aggressive a person seems in a  
3753 video. The most common measure of inter-annotator agreement is a  
3754 categorical measure called **Cohen’s  $\kappa$**  (“kappa”), for categorical agree-  
3755 ment, but you can use **intra-class correlation coefficients** (see the DEPTH  
3756 box below) for continuous data as well as many other measures.

### 🔍 DEPTH

#### *Reliability paradoxes!*

There’s a major issue with calculating reliabilities using the approaches we described here: because reliability is defined as a ratio of two measures of variation, it will always be relative to the variation in the sample. So if a sample has less variability, reliability will decrease!

One way to define reliability formally is by using the intra-class correla-

tion coefficient (ICC):

$$ICC = \frac{\sigma_b^2}{\sigma_w^2 + \sigma_b^2}$$

where  $\sigma_w^2$  is the within-subject variance in measurements and  $\sigma_b^2$  is the between-subject variance in the measurements. (The denominator of the ICC comes from partitioning the total observed variance  $\sigma_o^2$  in the reliability formula above.)

So now, instead of comparing variation to the mean, we're comparing variation on one dimension (between participants) to total variation (within and between participants). ICCs are tricky, and there are several different flavors available depending on the structure of your data and what you're trying to do with them. McGraw and Wong (1996) and Gwet (2014) provide extensive guidance on how to compute and interpret this statistic in different situations.

Let's think about the CDI data in our case study, which showed high reliability. Now imagine we restricted our sample to only change scores between 16–18-month-olds (our prior sample had 16–30-month-olds). Within this more restricted subset, overall vocabularies would be lower and more similar to one another, and so the average amount of change *within* a child ( $\sigma_w$ ) would be larger relative to the differences *between* children ( $\sigma_b$ ). That would make our reliability go *down*, even though we would be computing it on a subset of the exact same data.

That doesn't sound so bad. But we can construct a much more worrisome version of the same problem. Say we are very sloppy in our administra-

tion of the CDI and create lots of between-participants variability, perhaps by giving different instructions to different families. This practice will actually *increase* our estimate of split-half reliability (by increasing  $\sigma_b$ ). While the within-participant variability will remain the same, the between-participant variability will go up! You could call this a “reliability paradox”—sloppier data collection can actually lead to higher reliabilities.

We need to be sensitive to the sources of variability we’re quantifying reliability over—both the numerator and the denominator. If we’re computing split-half reliabilities, typically we’re looking at variability across test questions (from some question bank) vs across individuals (from some population). Both of these sampling decisions affect reliability—if the population is more variable *or* the questions are less variable, we’ll get higher reliability. In sum, *reliability is relative*: reliability measures depend on the circumstances in which they are computed.

3759

### 3760 8.1.1 Practical advice for computing reliability

3761 If you don’t know the reliability of your measures for an experiment,

3762 you risk wasting your and your participants’ time. Ignorance is not bliss.

3763 A higher reliability measure will lead to more precise measurements of

3764 a causal effect of interest and, hence, smaller required sample sizes.

3765 Test-retest reliability is generally the most conservative practical mea-

3766 sure of reliability. Test-retest estimates include not only measurement  
3767 error but also participants' state variation across different testing sessions  
3768 and variance due to differences between versions of your instrument.

3769 These real-world quantities are absent from internal reliability estimates,  
3770 which may make you erroneously think that there is more signal present  
3771 in your instrument than there is.<sup>8</sup> It's hard work to measure test-retest  
3772 reliability estimates, in part because you need two different versions of a  
3773 test (to avoid memory effects). If you plan on using an instrument more  
3774 than once or twice, though, it will likely be worthwhile!

3775 Finally, if you have multiple measurement items as part of your instru-  
3776 ment, make sure you evaluate how they contribute to the reliability  
3777 of the instrument. Perhaps you have several questions in a survey that  
3778 you'd like to use to measure the same construct; perhaps multiple exper-  
3779 imental vignettes that vary in content or difficulty. Some of these items  
3780 may not contribute to your instrument's reliability—and some may even  
3781 detract. At a bare minimum, you should always visualize the distribu-  
3782 tion of responses across items to scan for **floor and ceiling effects**—when  
3783 items always yield responses bunched at the bottom or top of the scale,  
3784 limiting their usefulness—and take a look at whether there are particu-  
3785 lar items on which items do not relate to the others.

3786 If you are thinking about developing an instrument that you use repeat-

<sup>8</sup> Even though  $\alpha$  is a theoretical lower bound on reliability, in practice, test-retest accuracy often ends up lower than  $\alpha$  because it incorporates all these other sources of variation.

3787 edly, it may be useful to use more sophisticated psychometric models  
3788 to estimate the dimensionality of responses on your instrument as well  
3789 as the properties of the individual items. If your items have binary an-  
3790 swers, like test questions, then **item response theory** is a good place to  
3791 start (Embreton and Reise 2013). If your items are more like ratings  
3792 on a continuous (interval or ratio) scale, then you may want to look at  
3793 factor analysis and related methods (Furr 2021).

### ⚠ ACCIDENT REPORT

#### *Wasted effort*

Low-reliability measures limit your ability to detect correlations between measurements. Mike spent several fruitless months in graduate school running dozens of participants through batteries of language processing tasks and correlating the results across tasks. Every time data collection finished, one or the other (spurious) correlation would show up in the data analysis. Something was always correlated with something else. Thankfully, he would always attempt to replicate the correlation in a new sample—and in that next dataset, the correlation we were trying to replicate would be null but another (again likely spurious) correlation would show up.

This exercise was a waste of time because most of the tasks were of such low reliability that, even had they been highly correlated with one another, relationship would have been almost impossible to detect without

a huge sample size. (It also would have been helpful if someone had mentioned multiplicity corrections [chapter 6] to him.)

One rule of thumb that's helpful for individual difference designs of this sort is that the maximal correlation that can be observed between two variables  $x$  and  $y$  is the square root of the product of their reliabilities:  $\sqrt{r_x r_y}$ . So if you have two measures that are reliable at 0.25, the maximal measured correlation between them is 0.25 as well! This kind of method is now frequently used in cognitive neuroscience (and other fields as well) to compute the so-called **noise ceiling** for a measure: the maximum amount of signal that in principle *could* be predicted (Lage-Castellanos et al. 2019). If your sample size is too small to detect correlations at the noise ceiling (see chapter 10), then the study is not worth doing.

3795

## 3796 8.2 Validity

3797 In chapter 2, we talked about the process of theory building as a process  
3798 of describing the relationships between constructs. But for the theory  
3799 to be tested, the constructs must be measured so that you can test the  
3800 relationships between them! Measurement and measure construction  
3801 is therefore intimately related to theory construction, and the notion of  
3802 validity is central.<sup>9</sup>

Shadish,

Cook, and Campbell 2002

3803 A valid instrument measures the construct of interest. In figure 8.1, in-

3804 validity is pictured as bias—the holes in the target are tightly grouped  
3805 but in the wrong place.<sup>10</sup> How can you tell if a measure is valid, given  
3806 that the construct of interest is unobserved? There is no single test of the  
3807 validity of a measure (Cronbach and Meehl 1955). Rather, the measure  
3808 is valid if there is evidence that fits into the broader theory as it relates  
3809 to the specific construct it is supposed to be measuring. For example, it  
3810 should be strongly related to other measures of the construct, but not as  
3811 related to measures of different constructs.

3812 How do you establish that a measure fits into the broader theory? Va-  
3813 lidity of a measure is typically established via an argument that calls on  
3814 different sources of support (Kane 1992). Here are some of the ways  
3815 that you might support the relationship between a measure and a con-  
3816 struct:

3817 – **Face validity:** The measure looks like the construct, such that in-  
3818 tuitively it is reasonable that it measures the construct. Face valid-  
3819 ity is a relatively weak source of evidence for validity, since it re-  
3820 lies primarily on pretheoretic intuitions rather than any quantita-  
3821 tive assessment. For example, reaction time is typically correlated  
3822 with intelligence test results (e.g., Jensen and Munro 1979) but  
3823 does not appear to be a face-valid measure of intelligence in that  
3824 simply being fast doesn't accord with our intuition about what it

<sup>10</sup> This metaphor is a good rough guide but it doesn't distinguish an instrument that is systematically biased (for example, by estimating scores too low for one group) and one that is invalid (because it measures the wrong construct).

means to be intelligent!

- **Ecological validity:** The measure relates to the context of people’s lives. For example, a rating of a child’s behavioral self-control in the classroom is a more ecologically valid measure of executive function than a reaction-time task administered in a lab context. Ecological validity arguments can be made on the basis of the experimental task, the stimuli, and the general setting of the experiment (Schmuckler 2001). Researchers differ in how much weight they assign to ecological validity based on their goals and their theoretical orientation.

- **Internal validity:** Usually used negatively. A “challenge to internal validity” is a description of a case where the measure is administered in such a way as to weaken the relationship between measure and construct. For example, if later items on a math test showed lower performance due to test-taker’s fatigue rather than lower knowledge of the concepts, the test might have an internal validity issue.<sup>11</sup>

- **Convergent validity:** The classic strategy for showing validity is to show that a measure relates (usually, correlates) with other putative measures of the same construct. When these relationships are measured concurrently, this is sometimes called **concurrent validity**. As we mentioned in chapter 2, self-reports of happiness

<sup>11</sup> Often this concept is described as being relevant to the validity of a *manipulation* also, for example, when the manipulation of the construct is confounded and some other psychological variable is manipulated as well. We discuss internal validity further in chapter 9.

3847 relate to independent ratings by friends and family, suggesting that  
3848 both measure the same underlying construct (Sandvik, Diener,  
3849 and Seidlitz 1993).<sup>12</sup>

- 3850 – **Predictive validity.** If the measure predicts other later measures  
3851 of the construct, or related outcomes that might be of broader  
3852 significance. Predictive validity is often used in lifespan and de-  
3853 velopmental studies where it is particularly prized for a measure  
3854 to be able to predict meaningful life outcomes such as educational  
3855 success in the future. For example, classroom self-control rat-  
3856 ings (among other measures) appear strongly predictive of later  
3857 life health and wealth outcomes (Moffitt et al. 2011).
- 3858 – **Divergent validity.** If the measure can be shown to be distinct  
3859 from measure(s) of a different construct, this evidence can help  
3860 establish that the measure is specifically linked to the target con-  
3861 struct. For example, measures of happiness (specifically, life satis-  
3862 faction) can be distinguished from measures of optimism as well as  
3863 both positive and negative affect, suggesting that these are distinct  
3864 constructs (Lucas, Diener, and Suh 1996).

<sup>12</sup> This idea of convergent validity re-  
lates to the idea of holism we described  
in chapter 2. A measure is valid if it  
relates to other valid measures, which  
themselves are only valid if the first one  
is! The measures are valid because the  
theory works, and the theory works be-  
cause the measures are valid. This circu-  
larity is a difficult but perhaps unavoid-  
able part of constructing psychological  
theories (see the DEPTH box above on  
the history of measurement). We don't  
ever have an objective starting point for  
the study of the human mind.

### 3865 8.2.1 Validity arguments in practice

3866 Let's take a look at how we might make an argument about the validity  
3867 of the CDI, the vocabulary instrument from our case study.

3868 First, the CDI is face valid—it is clearly about early language ability. In  
3869 contrast, even though a child's height would likely be correlated with  
3870 their early language ability, we should be skeptical of this measure due  
3871 to its lack of face validity. In addition, the CDI shows good convergent  
3872 and predictive validity. Concurrently, the CDI correlates well with evi-  
3873 dence from transcripts of children's actual speech and from standardized  
3874 language assessments (as discussed in the case study above). And predic-  
3875 tively, CDI scores at age two relate to reading scores during elementary  
3876 school (Marchman and Fernald 2008).

3877 On the other hand, users of the CDI must avoid challenges to the in-  
3878 ternal validity of the data they collect. For example, some CDI data  
3879 are compromised by confusing instructions or poor data collection pro-  
3880 cesses (Frank et al. 2021). Further, advocates and critics of the CDI  
3881 argue about its ecological validity. There is something quite ecologi-  
3882 cally valid about asking parents and caregivers—who are experts on their  
3883 own child—to report on their child's abilities. On the other hand, the  
3884 actual experience of filling out a structured form estimating language  
3885 ability might be more familiar to some families from higher education  
3886 backgrounds than for others from lower education backgrounds. Thus,  
3887 a critic could reasonably say that comparisons of CDI scores across so-  
3888 cioeconomic strata would be an invalid usage (Feldman et al. 2000).

3889 8.2.2 *Avoid questionable measurement practices!*

3890 Experimentalists sometimes have a tendency to make up ad hoc mea-  
3891 sures on the fly. It's fine to invent new measures, but the next step is  
3892 to think about what evidence there is that it's valid! Table 8.2 gives a  
3893 set of questions to guide thoughtful reporting of measurement practices  
3894 (adapted from [Flake and Fried 2020](#)).

Table 8.2  
Questions about measurement that every researcher should answer in their paper.  
Adapted from Flake and Fried (2020).

Question	Information to Report
What is your construct?	Define construct, describe theory and research.
What measure did you use to operationalize your construct?	Describe measure and justify operationalization.
Did you select your measure from the literature or create it from scratch?	Justify measure selection and review evidence on reliability and validity (or disclose the lack of such evidence).
Did you modify your measure during the process?	Describe and justify any modifications; note whether they occurred before or after data collection.

---

Question	Information to Report
How did you quantify your measure?	Describe decisions underlying the calculation of scores on the measure; note whether these were established before or after data collection and whether they are based on standards from previous literature.

---

3895 One big issue to be careful about is that researchers have been known to  
3896 modify their scales and their scale scoring practices (say, omitting items  
3897 from a survey or rescaling responses) after data collection. This kind of  
3898 post hoc alteration of the measurement instrument can sometimes be  
3899 justified by features of the data, but it can also look a lot like *p*-hacking!  
3900 If researchers modify their measurement strategy after seeing their data,  
3901 this decision needs to be disclosed, and it may undermine their statistical  
3902 inferences.

#### 💥 ACCIDENT REPORT

*Talk about flexible measurement!*

The competitive reaction time task (CRTT) is a lab-based measure of aggression. Participants are told that they are playing a reaction-time game against another player and are asked to set the parameters of a noise blast that will be played to their opponent. Unfortunately, in an analysis of

the literature using CRTT, Elson et al. (2014) found that different papers using the CRTT use dramatically different methods for scoring the task. Sometimes the analysis focused on the volume of the noise blast and sometimes it focused on the duration. Sometimes these scores were transformed (via logarithms) or thresholded. Sometimes they were combined into a single score. Elson was so worried by this flexibility that he created a website, <https://flexiblemeasures.com>, to document the variation he observed.

As of 2016, Elson had found 130 papers using the CRTT. And across these papers, he documented an astonishing 157 quantification strategies. One paper reported ten different strategies for extracting numbers from this measure! More worrisome still, Elson and colleagues found that when they tried out some of these strategies on their own data, different strategies led to very different effect sizes and levels of statistical significance. They could effectively make a finding appear bigger or smaller depending on which scoring they chose.

Triangulating a construct through multiple prespecified measurements can be a good thing. But the issue with the CRTT analysis was that changes in the measurement strategy appeared to be made in a post hoc, data-driven way so as to maximize the significance of the experimental manipulation (just like the *p*-hacking we discussed in chapters 3 and 6).

This examination of the use of the CRTT measure has several implications. First, and most troublingly, there may have been undisclosed flex-

ability in the analysis of CRTT data across the literature, with investigators taking advantage of the lack of standardization to try many different analysis variants and report the one most favorable to their own hypothesis. Second, it is unknown which quantification of CRTT behavior is in fact most reliable and valid. Since some of these variants are presumably better than others, researchers are effectively “leaving money on the table” by using suboptimal quantifications. As a consequence, if researchers adopt the CRTT, they find much less guidance from the literature on what quantification to adopt.

3905

### 3906 8.3 How to select a good measure?

3907 Ideally you want a measure that is reliable and valid. How do you get  
3908 one? An important first principle is to use a preexisting measure. Per-  
3909 haps someone else has done the hard work of compiling evidence on  
3910 reliability and validity, and in that case, you will most likely want to pig-  
3911 gyback on that work. Standardized measures are typically broad in their  
3912 application, and so the tendency can be to discard these because they are  
3913 not tailored for our studies specifically. But the benefits of a standar-  
3914 ized measure are substantial. Not only can you justify the measure using  
3915 the prior literature, but you also have an important index of population  
3916 variability by comparing absolute scores to other reports.<sup>13</sup>

3917 If you don't use someone else's measure, you'll need to make one up  
3918 yourself. Most experimenters go down this route at some point, but if  
3919 you do, remember that you will need to figure out how to estimate its  
3920 reliability and also how to make an argument for its validity!

3921 We can assign numbers to almost anything people do. We could run  
3922 an experiment on children's exploratory play and count the number of  
3923 times they interact with another child (Ross and Lollis 1989), or run an  
3924 experiment on aggression where we quantify the amount of hot sauce  
3925 participants serve (Lieberman et al. 1999). Yet, most of the time we  
3926 choose from a relatively small set of operational variables: asking survey  
3927 questions, collecting choices and reaction times, and measuring physio-  
3928 logical variables like eye movements. Besides following these conven-  
3929 tions, how do we choose the right measurement type for a particular  
3930 experiment?

3931 There's no hard and fast rule about what aspect of behavior to measure,  
3932 but here we will focus on two dimensions that can help us organize the  
3933 broad space of possible measure targets.<sup>14</sup> The first of these is the contin-  
3934 uum between simple and complex behaviors. The second is the focus  
3935 on explicit, voluntary behaviors vs implicit or involuntary behaviors.

<sup>14</sup> Some authors differentiate between "self-report" and "observational" measures. This distinction seems simple on its face but actually gets kind of complicated. Is a facial expression a "self-report"? Language is not the only way that people communicate with one another—many actions are intended to be communicative (Shafto, Goodman, and Frank 2012).

3936 8.3.1 *Simple vs complex behaviors*

3937 Figure 8.5 shows a continuum between simple and complex behaviors.

3938 The simplest measurable behaviors tend to be button presses, for exam-

3939 ple:

- 3940 – pressing a key to advance to the next word in a word-by-word
- 3941 self-paced reading study;
- 3942 – selecting “yes” or “no” in a lexical decision task; and
- 3943 – making a forced choice between different alternatives to indicate which has been seen before.

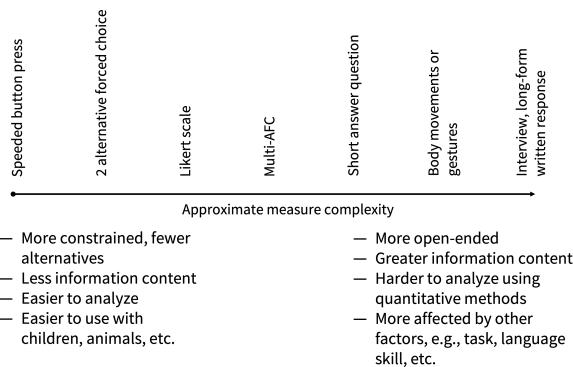


Figure 8.5

Often choosing a measure can be consolidated into a choice along a continuum from simple measures that provide a small amount of information but are quick and easy to repeat and those that provide much richer information but require more time.

3945 These specific measures—and many more like them—are the bread and

3946 butter of many cognitive psychology studies. Because they are quick

3947 and easy to explain, these tasks can be repeated over many trials. They

3948 can also be executed with a wider variety of populations including with

3949 young children and sometimes even with nonhuman animals with ap-

3950 propriate adaptation. (A further benefit of these paradigms is that they

3951 can yield useful reaction time data, which we discuss further below.)

3952 In contrast, a huge range of complex behaviors have been studied by

3953 psychologists, including:

3954 – open-ended verbal interviews;

3955 – written expression, for example, via handwriting or writing style;

3956 – body movements, including gestures, walking, or dance; and

3957 – drawing or artifact building.

3958 There are many reasons to study these kinds of behaviors. First, the

3959 behaviors themselves may be examples of tasks of interest (e.g., studies

3960 of drawing that seek to understand the origins of artistic expression).

3961 Or, the behavior may stand in for other even more complex behaviors

3962 of interest, as in studies of typing that use this behavior as a proxy for

3963 lexical knowledge (Rumelhart and Norman 1982).

3964 Complex behaviors typically afford a huge variety of different measure-

3965 ment strategies. So any experiment that uses a particular measurement

3966 of a complex behavior will typically need to do significant work up

3967 front to justify the choice of that measurement strategy—for example,

3968 how to quantify dances or gestures or typing errors—and provide some

3969 assurance about its reliability. Further, it is often much more difficult

3970 to have a participant repeat a complex behavior many times under the

3971 same conditions. Imagine asking someone to draw hundreds of sketches  
3972 as opposed to pressing a key hundreds of times! Thus, the choice of a  
3973 complex behavior is often a choice to forego a large number of simple  
3974 trials for a small number of more complex trials.

3975 Complex behaviors can be especially useful to study either at the be-  
3976 ginning or the end of a set of experiments. At the beginning of a set of  
3977 experiments, they can provide inspiration about the richness of the tar-  
3978 get behavior and insight into the many factors that influence it. And at  
3979 the end, they can provide an ecologically valid measure to complement  
3980 a reliable but more artificial lab-based behavior.

3981 The more complex the behavior, however, the more it will vary across  
3982 individuals and the more environmental and situational factors will af-  
3983 fect it. These can be important parts of the phenomenon, but they can  
3984 also be nuisances that are difficult to get under experimental control.<sup>15</sup>  
3985 Simple measures are typically easier to use and, hence, easier to deploy  
3986 repeatedly in a set of experiments where you iterate your manipulation  
3987 to test a causal theory.

### 3988 8.3.2 *Implicit vs explicit behaviors*

3989 A second important dimension of organization for measures is the dif-  
3990 ference between implicit and explicit measures. An explicit measure

<sup>15</sup> When they are not designed with care, complex, open-ended behaviors such as verbal interviews can be especially affected by the experimental biases that we will describe in chapter 9, including, for example, **demand characteristics**, in which participants say what they think experimenters want to hear. Qualitative interview methods can be incredibly powerful as a method in their own right, but they should be deployed with care as measures for an experimental intervention.

3991 provides a measurement of a behavior that a participant has conscious  
3992 awareness of—for example, the answer to a question. In contrast, im-  
3993 plicit measures provide measurements of psychological processes that  
3994 participants are unable to report (or occasionally, unwilling to).<sup>16</sup> Im-  
3995 plicit measures, especially reaction time, have long been argued to re-  
3996 flect internal psychological processes (Donders 1969 [1868]). They also  
3997 have been proposed as measures of qualities such as racial bias that partic-  
3998 ipants may have motivation not to disclose (Greenwald, McGhee, and  
3999 Schwartz 1998). There are also of course a host of physiological mea-  
4000 surements available. Some of these measure eye movements, heart rate,  
4001 or skin conductance, which can be linked to aspects of cognitive pro-  
4002 cess. Others reflect underlying brain activity via the signals associated  
4003 with MRI (magnetic resonance imaging), MEG (magnetoencephalogra-  
4004 phy), NIRS (near-infrared spectroscopy), and EEG (electroencephalo-  
4005 gram) measurements. These methods are outside the scope of this book,  
4006 though we note that the measurement concerns we discuss here defi-  
4007 nitely apply (e.g., Zuo, Xu, and Milham 2019).

4008 Many tasks produce both accuracy and reaction time data. Often these  
4009 trade off with one another in a classic **speed-accuracy trade-off**: the  
4010 faster participants respond, the less accurate they are. For example, to  
4011 investigate racial bias in policing, Payne (2001) showed US college stu-  
4012 dents a series of pictures of tools and guns, proceeded by a prime of

<sup>16</sup> Implicit/explicit is likely more of a continuum, but one cut-point is whether the participants' behavior is considered intentional: that is, participants *intend* to press a key to register a decision, but they likely do not intend to react in 300 as opposed to 350 milliseconds due to having seen a prime.

4013 either a White face or a Black face. In a first study, participants were  
4014 faster to identify weapons when primed by a Black face but had similar  
4015 accuracies. A second study added a response deadline to speed up judg-  
4016 ments: this manipulation resulted in equal reaction times across condi-  
4017 tions but greater errors in weapon identification after Black faces. These  
4018 studies likely revealed the same phenomenon—some sort of bias to asso-  
4019 ciate Black faces with weapons—but the design of the task moved par-  
4020 ticipants along a speed-accuracy trade-off, yielding effects on different  
4021 measures.<sup>17</sup>

4022 Simple, explicit behaviors are often a good starting point. Work using  
4023 these measures—often the least ecologically valid—can be enriched with  
4024 implicit measures or measurements of more complex behaviors.

#### 4025 *8.4 The temptation to measure lots of things*

4026 If one measure is good, shouldn't two be better? Many experimenters  
4027 add multiple measurements to their experiments, reasoning that more  
4028 data is better than less. But that's not always true!

4029 Deciding whether to include multiple measures is an aesthetic and prac-  
4030 tical issue as well as a scientific one. Throughout this book, we have  
4031 been advocating for a viewpoint in which experiments should be as

<sup>17</sup> One way of describing the information processing underlying this trade-off is given by drift diffusion models (DDMs), which allow joint analysis of accuracy and reaction time (Voss, Nagerl, and Lerche 2013). Used appropriately, DDMs can provide a way to remove speed-accuracy trade-offs and extract more reliable signals from tasks where accuracy and reaction time are both measured (see Johnson et al. 2017 for an example of a DDM on a weapon-decision task).

4032 simple as possible. For us, the best experiment is one that shows that  
4033 a simple and valid manipulation affects a single measure that is reliable  
4034 and valid.<sup>18</sup> If you are tempted to include more than one measure, see  
4035 if we can talk you out of it.<sup>19</sup>

4036 First, make sure that including more measures doesn't compromise each  
4037 individual measure. This can happen via fatigue or carryover effects.  
4038 For example, if a brief attitude manipulation is followed by multiple  
4039 questionnaire measures, it is a good bet that there is likely to be "fade-  
4040 out" of the effect over time, so it won't have the same effect on the  
4041 first questionnaire as the last one. Further, even if a manipulation has  
4042 a long duration effect on participants, survey fatigue may lead to less  
4043 meaningful responses to later questions (Herzog and Bachman 1981).

4044 Second, consider whether you have a strong prediction for each measure,  
4045 or whether you're just looking for more ways to see an effect of your  
4046 manipulation. As discussed in chapter 2, we think of an experiment as  
4047 a "bet." The more measures you add, the more bets you are making  
4048 but the less value you are putting on each. In essence, you are "hedging  
4049 your bets," and so the success of any one bet is less convincing.

4050 Third, if you include multiple measures in your experiment, you need  
4051 to think about how you will interpret inconsistent results. Imagine you  
4052 have experimental participants engage in a brief written reflection that

<sup>18</sup> In an entertaining article called "Things I Have Learned (So Far)," Cohen (1990) quips that he leans so far in the direction of large numbers of observations and small numbers of measures that some students think his perfect study has 10,000 participants and *no* measures.

<sup>19</sup> As usual, we want to qualify that we are only talking about randomized experiments here! In observational studies, often the point is to measure the associations between multiple measures, so you typically *have* to include more than one. Additionally, some of the authors of this book have advocated for measuring multiple outcomes in longitudinal observational studies, which could reduce investigator bias, encourage reporting null effects, enable comparison of effect sizes, and improve research efficiency (VanderWeele, Mathur, and Chen 2020). We've also done plenty of descriptive studies—these can be very valuable. In a descriptive context, often the goal is to include as many measures as possible so as to have a holistic picture of the phenomenon of interest.

4053 is hypothesized to affect a construct (vs a control writing exercise, say  
4054 listing meals). If you include two measures of the construct of inter-  
4055 est and one shows a larger effect, what will you conclude? It may be  
4056 tempting to assume that the one that shows a larger effect is the “better  
4057 measure,” but the logic is circular—it’s only better if the manipulation  
4058 affected the construct of interest, which is what you were testing in the  
4059 first place! Including multiple measures because you’re uncertain which  
4060 one is more related to the construct indulges in this circular logic, since  
4061 the experiment often can’t resolve the situation. A much better move  
4062 in this case is to do a preliminary study of the reliability and validity  
4063 of the two measures so as to be able to select one as the experiment’s  
4064 primary endpoint.<sup>20</sup>

4065 Finally, if you do include multiple measures, selective reporting of sig-  
4066 nificant or hypothesis-aligned measures becomes a real risk. For this rea-  
4067 son, preregistration and transparent reporting of all measures becomes  
4068 even more important.

4069 There are some cases where more measures are better. The more expen-  
4070 sive the experiment, the less likely it is to be repeated to gather a new  
4071 measurement of the effects of the same manipulation. Thus, larger stud-  
4072 ies present a stronger rationale for including multiple measures. Clinical  
4073 trials often involve interventions that can have effects on many different

<sup>20</sup> One caveat to this argument is that it can sometimes be useful to examine the effects of a manipulation on different measures because the measures are important. For example, you might be interested in whether an educational intervention increased grades *and* decreased dropout rates. Both outcome measures are important, and so it is useful to include both in your study.

4074 measures; imagine a cancer treatment that might affect mortality rates,  
4075 quality of life, tumor growth rates, and other measures. Further, such  
4076 trials are extremely expensive and difficult to repeat. Thus, there is a  
4077 good reason for including more measures in such studies.

 DEPTH

### *Survey measures*

Sometimes the easiest way to elicit information from participants is simply to ask. Surveys are an important part of experimental measurement, so we'll share a few best practices, primarily derived from Krosnick and Presser (2010).

Treat survey questions as a conversation. The easier your items are to understand, the better. Don't repeat variations on the same question unless you want different answers! Try to make the order reasonable, for example, by grouping together questions about the same topic and moving from more general to more specific questions. The more you include "tricky" items, the more you invite tricky answers to straightforward questions. One specific kind of tricky questions are "check" questions that evaluate participant compliance. We'll talk more in chapter 12 about various ways of evaluating compliance and their strengths and weaknesses.

Open-ended survey questions can be quite rich and informative, especially when an appropriate coding (classification) scheme is developed in advance and responses are categorized into a relatively small number of types. On the other hand, they present practical obstacles because they

require coding (often by multiple coders to ensure reliability of the coding). Further, they tend to yield nominal data, which are often less useful for quantitative theorizing. Open-ended questions are a useful tool to add nuance and color to the interpretation of an experiment.

One common mistake that survey developers make is trying to put too much into one question. Imagine asking a restaurant-goer for a numerical ranking on the question, “How do you like our food and service?” What if they loved the food but hated the service, or vice versa? Would they choose an intermediate option? Items that ask about more than one thing at once are known as **double-barreled** questions. They can confuse and frustrate participants as well as lead to uninterpretable data.

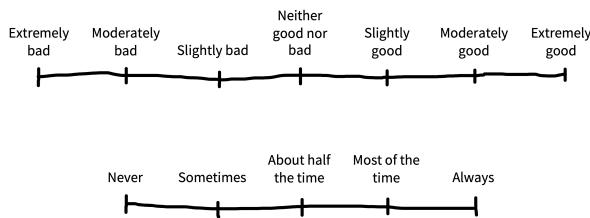


Figure 8.6  
Likert scales based on survey best practices: a bipolar opinion scale with seven points and a unipolar frequency scale with five points. Both have all points labeled.

Especially given their ubiquity in commercial survey research, **Likert scales**—scales with a fixed number of ordered, numerical response options—are a simple and conventional way of gathering data on attitude and judgment questions (figure 8.6). Bipolar scales are those in which the end points represent opposites: for example the continuum between “strongly dislike” and “strongly like.” Unipolar scales have one neutral endpoint, like the continuum between “no pain” and “very intense pain.”

Survey methods research suggests that reliability is maximized when bipolar scales have seven points and unipolar scales have five. Labeling every point on the scale with verbal labels is preferable to labeling only the endpoints.

One important question is whether to treat data from Likert scales as ordinal or interval. It's extremely common (and convenient) to make the assumption that Likert ratings are interval, allowing the use of standard statistical tools like means, standard deviations, and linear regression. The risk in this practice comes from the possibility that scale items are not evenly spaced—for example, on a scale labeled “never,” “seldom,” “occasionally,” “often,” and “always,” the distance from “often” to “always” may be larger than the distance from “seldom” to “occasionally.”

In practice, you can choose to use regression variants that are appropriate, for example, ordinal logistic regression and its variants, or they can attempt to assess and mitigate the risks of treating the data as interval. If you choose the second option, it's definitely a good idea to look carefully at the raw distributions for individual items to see if their distribution appears approximately normal (see chapter 15).

Recently some researchers have begun to use “visual analog scales” (or sliders) as a solution. We don't recommend these—the distribution of the resulting data is often anchored at the starting point or endpoints (Matejka et al. 2016), and a meta-analysis shows they're a lot lower than Likert scales in reliability (Krosnick and Presser 2010).

It rarely helps matters to add an “I don’t know” or “other” option to survey questions. These are some of a variety of practices that encourage **satisficing**, where survey takers give answers that are good enough but don’t reflect substantial thought about the question. Another behavior that results from satisficing is “straight-lining”—that is, picking the same option for every question. In general, the best way to prevent straight-lining is to make surveys relatively short, engaging, and well-compensated. The practice of “reverse coding” to make the expected answers to some questions more negative can block straight-lining, but at the cost of making items more confusing. Some obvious formatting options can reduce straight-lining as well: for example, placing scales further apart or on subsequent (web) pages.

In sum, survey questions can be a helpful tool for eliciting graded judgments about explicit questions. The best way to execute them well is to try and make them as clear and easy to answer as possible.

4081

## 4082 *8.5 Chapter summary: Measurement*

4083 In olden times, all the psychologists went to the same conferences and  
4084 worried about the same things. But then a split formed between differ-  
4085 ent groups. Educational psychologists and psychometricians thought  
4086 a lot about how different problems on tests had different measurement  
4087 properties. They began exploring how to select good and bad items, and

4088 how to figure out people's ability abstracted away from specific items.

4089 This research led to a profusion of interesting ideas about measurement,

4090 but these ideas rarely percolated into day-to-day practice in other areas

4091 of psychology. For example, cognitive psychologists collected lots of tri-

4092 als and measured quantities of interest with high precision, but they wor-

4093 ried less about measurement validity. Social psychologists spent more

4094 time worrying about issues of ecological validity in their experiments,

4095 but they often used ad hoc scales with poor psychometric properties.

4096 These sociological differences between fields have led to an unfortunate

4097 divergence, where experimentalists often don't recognize the value of

4098 the conceptual tools developed to aid measurement, and so fail to reason

4099 about the reliability and validity of their measures in ways that can help

4100 them make better inferences. As we said in our discussion of reliability,

4101 ignorance is not bliss. Much better to think these choices through!



### DISCUSSION QUESTIONS

1. Let's go back to our example on the relationship between money and happiness from chapter 1. How many different kinds of measures of happiness can you come up with? Make a list with at least five.
2. Choose one of your measures of happiness and come up with a validation strategy for it, making reference to at least three different types of validity. What data collection would this validation effort require?



## READINGS

- A classic textbook on psychometrics that introduces the concepts of reliability and validity in a simple and readable way: Furr, R. Michael (2021). *Psychometrics: An Introduction*. SAGE publications.
- A great primer on questionnaire design: Krosnick, Jon A. (2018). “Improving Question Design to Maximize Reliability and Validity.” In *The Palgrave Handbook of Survey Research*, edited by David L. Vannette and Jon A. Krosnick, 95–101. Springer. [https://doi.org/10.1007/978-3-319-54395-6\\_13](https://doi.org/10.1007/978-3-319-54395-6_13).
- Introduction to general issues in measurement and why they shouldn’t be ignored: Flake, Jessica Kay, and Eiko I. Fried (2020). “Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them.” *Advances in Methods and Practices in Psychological Science* 3 (4): 456–465. <https://doi.org/10.1177/2515245920952393>.
- An accessible popular book on scientific measurement: Vincent, James (2022). *Beyond Measure: The Hidden History of Measurement from Cubits to Quantum Constants*. Faber & Faber.

## 4104 References

- Bornstein, Marc H., and O. Maurice Haynes. 1998. "Vocabulary Competence in Early Childhood: Measurement, Latent Construct, and Predictive Validity." *Child Development* 69 (3): 654–671.
- Brandmaier, Andreas M., Elisabeth Wenger, Nils C. Bodammer, Simone Kühn, Naftali Raz, and Ulman Lindenberger. 2018. "Assessing Reliability in Neuroimaging Research through Intra-Class Effect Decomposition 4105 (ICED)." *Elife* 7:e35718.
- Campbell, Norman Robert. 1928. *An Account of the Principles of Measurement and Calculation*. Longmans, Green & Company.
- Cattel, J. McK. 1890. "Mental Tests and Measurements." *Mind* 15:373–380.
- Cattell, Raymond B. 1962. "The Relational Simplex Theory of Equal Interval and Absolute Scaling." *Acta Psychologica* 20:139–158.
- Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.
- Cohen, Jacob. 1990. "Things I Have Learned (So Far)." *American Psychologist* 45:1304–1312.
- Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4): 281–302.
- Darrigol, Olivier. 2003. "Number and Measure: Hermann von Helmholtz at the Crossroads of Mathematics, Physics, and Psychology." *Studies in History and Philosophy of Science Part A* 34 (3): 515–573.
- Donders, Franciscus Cornelis. 1969 [1868]. "On the Speed of Mental Processes." Translated by W. G. Koster. *Acta Psychologica* 30 (1969): 412–431.

Elson, Malte, M. Rohangis Mohseni, Johannes Breuer, Michael Scharkow, and Thorsten Quandt. 2014. “Press CRTT to Measure Aggressive Behavior: The Unstandardized Use of the Competitive Reaction Time Task in Aggression Research.” *Psychological Assessment* 26 (2): 419–432. <https://doi.org/10.1037/a0035569>.

Embretson, Susan E., and Steven P. Reise. 2013. *Item Response Theory*. Psychology Press.

Fechner, Gustav Theodor. 1987 [1887]. “My Own Viewpoint on Mental Measurement (1887).” Translated by Eckart Scheerer. *Psychological Research* 49 (4): 213–219.

Fechner, Gustav Theodor. 1860. *Elemente Der Psychophysik*. Vol. 2. Breitkopf u. Härtel.

Feldman, Heidi M., Christine A. Dollaghan, Thomas F. Campbell, Marcia Kurs-Lasky, Janine E. Janosky, and Jack L. Paradise. 2000. “Measurement Properties of the MacArthur Communicative Development Inventories at Ages One and Two Years.” *Child Development* 71 (2): 310–322.

Ferguson, A., C. S. Myers, R. J. Bartlett, H. Banister, F. C. Bartlett, Brown W., and W. S. Tucker. 1940. “Quantitative Estimates of Sensory Events, Final Report.” *Report of the British Association for the Advancement of Science*, 331–349.

Flake, Jessica Kay, and Eiko I. Fried. 2020. “Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them.” *Advances in Methods and Practices in Psychological Science* 3 (4): 456–465.

Frank, Michael C., Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2021. *Variability and Consistency in Early Language Learning: The Word-*

- bank Project. MIT Press.
- Furr, R. Michael. 2021. *Psychometrics: An Introduction*. SAGE publications.
- Giere, Ronald N. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71 (5): 742–752. <https://doi.org/10.1086/425063>.
- Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74 (6): 1464.
- Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics.
- Hardcastle, Gary L. 1995. "S. S. Stevens and the Origins of Operationism." *Philosophy of Science* 62 (3): 404–424.
- Heidelberger, Michael. 2004. *Nature from Within: Gustav Theodor Fechner and His Psychophysical Worldview*. University of Pittsburgh Press.
- Herzog, A. Regula, and Jerald G. Bachman. 1981. "Effects of Questionnaire Length on Response Quality." *Public Opinion Quarterly* 45 (4): 549–559.
- Jensen, Arthur R., and Ella Munro. 1979. "Reaction Time, Movement Time, and Intelligence." *Intelligence* 3 (2): 121–126.
- Johnson, David J., Christopher J. Hopwood, Joseph Cesario, and Timothy J. Pleskac. 2017. "Advancing Research on Cognitive Processes in Social and Personality Psychology: A Hierarchical Drift Diffusion Model Primer." *Social Psychological and Personality Science* 8 (4): 413–423.
- Kane, Michael T. 1992. "An Argument-Based Approach to Validity." *Psychological Bulletin* 112 (3): 527–535.
- Kisch, B. 1965. *Scales and Weights: A Historical Outline*. Yale Studies in the History of Science and Medicine. Yale University Press.

Krantz, David H., Robert Duncan Luce, Patrick Suppes, and Amos Tversky.

1971. *Foundations of Measurement I: Additive and Polynomial Representations*.  
Courier Corporation.

Krosnick, Jon A. 2018. “Improving Question Design to Maximize Reliability and Validity.” In *The Palgrave Handbook of Survey Research*, edited by David L. Vannette and Jon A. Krosnick, 95–101. Springer.

Krosnick, Jon A., and Stanley Presser. 2010. “Question and Questionnaire Design.” *Handbook of Survey Research*, 263.

Lage-Castellanos, Agustin, Giancarlo Valente, Elia Formisano, and Federico De Martino. 2019. “Methods for Computing the Maximum Performance of Computational Models of fMRI Responses.” *PLoS Computational Biology* 15 (3): e1006397.

Lieberman, Joel D., Sheldon Solomon, Jeff Greenberg, and Holly A. McGregor. 1999. “A Hot New Way to Measure Aggression: Hot Sauce Allocation.” *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 25 (5): 331–348.

Lucas, Richard E., Ed Diener, and Eunkook Suh. 1996. “Discriminant Validity of Well-Being Measures.” *Journal of Personality and Social Psychology* 71 (3): 616–628.

Luce, Robert D., David H. Krantz, Patrick Suppes, and Amos Tversky. 1990. *Foundations of Measurement III: Representation, Axiomatization, and Invariance*. Courier Corporation.

Luce, Robert D., and John W. Tukey. 1964. “Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement.” *Journal of Mathematical Psychology* 1 (1): 1–27.

- Marchman, Virginia A., and Anne Fernald. 2008. "Speed of Word Recognition and Vocabulary Knowledge in Infancy Predict Cognitive and Language Outcomes in Later Childhood." *Developmental Science* 11 (3): F9–F16.
- Matejka, Justin, Michael Glueck, Tovi Grossman, and George Fitzmaurice. 2016. "The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5421–5432.
- Maul, Andrew, David Torres Irribarra, and Mark Wilson. 2016. "On the Philosophical Foundations of Psychological Measurement." *Measurement* 79:311–320.
- McGraw, Kenneth O., and Seok P. Wong. 1996. "Forming Inferences about Some Intraclass Correlation Coefficients." *Psychological Methods* 1 (1): 30–46.
- Michell, Joel. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. Vol. 53. Cambridge University Press.
- Moffitt, Terrie E., Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J. Hancox, Honalee Harrington, Renate Houts, et al. 2011. "A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety." *Proceedings of the National Academy of Sciences* 108 (7): 2693–2698.
- Moscati, Ivan. 2018. *Measuring Utility: From the Marginal Revolution to Behavioral Economics*. Oxford University Press.
- Narens, Louis, and Robert Duncan Luce. 1986. "Measurement: The Theory of Numerical Assignments." *Psychological Bulletin* 99 (2): 166–180.
- Payne, B. Keith. 2001. "Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon." *Journal of Personality*

- and *Social Psychology* 81 (2): 181–192.
- Putnam, Hilary. 2000. *The Threefold Cord: Mind, Body, and World*. Columbia University Press.
- Ross, Hildy S., and Susan P. Lollis. 1989. “A Social Relations Analysis of Toddler Peer Relationships.” *Child Development* 60 (5): 1082–1091.
- Rumelhart, David E., and Donald A. Norman. 1982. “Simulating a Skilled Typist: A Study of Skilled Cognitive-Motor Performance.” *Cognitive Science* 6 (1): 1–36.
- Sandvik, Ed, Ed Diener, and Larry Seidlitz. 1993. “Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures.” *Journal of Personality* 61 (3): 317–342.
- Schmuckler, Mark A. 2001. “What Is Ecological Validity? A Dimensional Analysis.” *Infancy* 2 (4): 419–436.
- Shadish, William, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.
- Shafto, Patrick, Noah D. Goodman, and Michael C. Frank. 2012. “Learning from Others: The Consequences of Psychological Reasoning for Human Learning.” *Perspectives on Psychological Science* 7 (4): 341–351.
- Sijtsma, Klaas. 2009. “On the Use, the Misuse, and the Very Limited Usefulness of Cronbach’s Alpha.” *Psychometrika* 74 (1): 107–120.
- Stevens, S. S. 1946. “On the Theory of Scales of Measurement.” *Science* 103 (2684): 677–680.
- Suppes, Patrick, David H. Krantz, Robert Duncan Luce, and Amos Tversky. 1989. *Foundations of Measurement II: Geometrical, Threshold, and Probabilistic*

- Representations*. Courier Corporation.
- Tal, Eran. 2020. “Measurement in Science.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. <https://plato.stanford.edu/archives/fall2020/entries/measurement-science/>.
- VanderWeele, Tyler J., Maya B. Mathur, and Ying Chen. 2020. “Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies.” *Statistical Science* 35 (3): 437–466.
- Vincent, James. 2022. *Beyond Measure: The Hidden History of Measurement*. Faber & Faber.
- Voss, Andreas, Markus Nagler, and Veronika Lerche. 2013. “Diffusion Models in Experimental Psychology.” *Experimental Psychology* 60 (6): 385–402. <https://doi.org/10.1027/1618-3169/a000218>.
- Woodworth, R. S. 1938. *Experimental Psychology*. Holt.
- Zuo, Xi-Nian, Ting Xu, and Michael Peter Milham. 2019. “Harnessing Reliability for Neuroscience Research.” *Nature Human Behaviour* 3 (8): 768–771. <https://doi.org/10.1038/s41562-019-0655-x>.

## <sup>4113</sup> 9 DESIGN



### LEARNING GOALS

- Describe key elements of experimental design
- Define randomization and counterbalancing strategies for removing confounds
- Discuss strategies to design experiments that are appropriate to the populations of interest

<sup>4114</sup>

<sup>4115</sup> The key thesis of our book is that experiments should be designed to  
<sup>4116</sup> yield precise and unbiased measurements of a causal effect. But the  
<sup>4117</sup> causal effect of what? The manipulation! In an experiment we manipu-  
<sup>4118</sup> late (intervene on) some aspect of the world and measure the effects of  
<sup>4119</sup> that manipulation. We then compare that measurement to a situation  
<sup>4120</sup> where the intervention has not occurred.

<sup>4121</sup> We refer to different intervention states as **conditions** of the experiment.  
<sup>4122</sup> The most common experimental design is the comparison between a

4123 control condition, in which the intervention is not performed, and an  
4124 **experimental** (sometimes called **treatment**) condition in which the in-  
4125 tervention is performed.

4126 But many other experimental designs are possible. In more complex ex-  
4127 periments, manipulations along different dimensions (sometimes called  
4128 **factors** in this context) can be combined. In the first part of the chapter,  
4129 we'll introduce some common experimental designs and the vocabu-  
4130 lary for describing them. Our focus here is in identifying designs that  
4131 maximize MEASUREMENT PRECISION.

4132 A good experimental measure must be a valid measure of the construct  
4133 of interest. The same is true for a manipulation—it must validly relate to  
4134 the causal effect of interest. In the second part of the chapter, we'll dis-  
4135 cuss issues of **manipulation validity**, including both issues of ecological  
4136 validity and **confounding**. We'll talk about how practices like **random-**  
4137 **ization** and **counterbalancing** can help remove nuisance confounds, an  
4138 important part of **BIAS REDUCTION** for experimental designs.<sup>1</sup>

4139 To preview our general take-home points from this chapter: we think  
4140 that your default experiment should manipulate one or two factors—  
4141 usually not more—and should manipulate those factors continuously  
4142 and within participants. Although such designs are not always possible,

<sup>1</sup> This section will draw on our introduction to causal inference in chapter 1, so if you haven't read that, now's the time.

<sup>4143</sup> they are typically the most likely to yield precise estimates of a particu-  
<sup>4144</sup> lar effect that can be used to constrain future theorizing. We'll start by  
<sup>4145</sup> considering a case study in which a subtle confound led to difficulties  
<sup>4146</sup> interpreting an experimental result.



### CASE STUDY

#### *Automatic theory of mind?*

In an early version of our course, student Desmond Ong set out to replicate a thought-provoking finding: both infants and adults seemed to show evidence of tracking other agents' belief state, even when it was irrelevant to the task at hand (Kovács, Téglás, and Endress 2010). In the paradigm, an animated Smurf character would watch as a self-propelled ball came in and out from behind a screen. At the end of the video, the screen would swing down and the participant had to respond whether the ball was present or absent. Reaction time for this decision was the key dependent variable.

The experimental design investigated two factors: whether the participant believed the ball was present or absent ( $P+/P-$ ) and whether the animated agent *would have believed* the ball was present or absent ( $A+/A-$ ) based on what it saw. The result was four conditions:  $P+/A+$ ,  $P+/A-$ ,  $P-/A+$ , and  $P-/A-$ . (We could call this a **fully-crossed** design because each level of one factor was presented with each level of the other.)

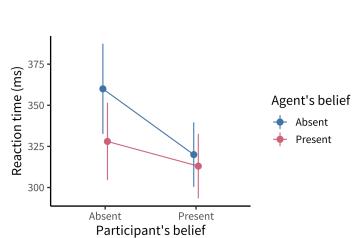


Figure 9.1

Original data from Kovács, Téglás, and Endress (2010). Error bars show 95% confidence intervals. Based on Phillips et al. (2015).

Both the original experiments and the replication that Desmond ran showed a significant effect of the agent's beliefs on participants' reaction times, suggesting that what the—totally irrelevant—agent thought about the ball was leading them to react more or less quickly to the presence of the ball. Figure 9.1 shows the original data ( $N = 24$ ). But, although both studies showed an effect of agent belief, the replication and several variations also showed a crossover interaction of participant and agent belief. The participants were slower when the agents *and* the participants believed that the ball was behind the screen (figure 9.2). That finding wasn't consistent with the theory that tracking inconsistent beliefs slowed down reaction times. If participants were tracking their own beliefs about the ball *and* the agent's, they should have been fastest in the P+/A+ condition, not slower.

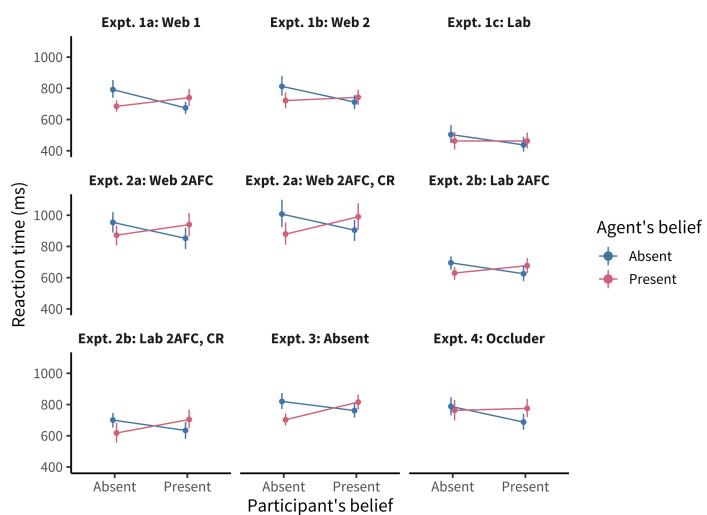


Figure 9.2

Data from a series of replications of Kovács, Téglás, and Endress (2010), including versions on the web (experiments 1a and 1b) and in lab (experiment 1c), as well as several variations on the format of responding (experiments 2 and 3; 2AFC = two alternative forced choice) and an experiment where a large wall kept the agent from seeing the ball at all (experiment 4). “Hits” and “CRs” panels refer to different subsets of trials where participants responded “present” when the ball was present and “absent” when the ball was absent. Error bars are 95% confidence intervals. Based on Phillips et al. (2015).

A collaborative team working on this paradigm identified a key issue (Phillips et al. 2015). There was a **confound** in the experimental design—another factor that varied across conditions besides the target factors. In other words, something was changing between conditions other than the agent’s and participant’s belief states. The confound was an attention check (discussed further in chapter 12): participants had to press a key when the agent left the scene to show that they were paying attention. This attention check appeared a few seconds later in the videos for the P+/A+ and P-/A- trials—the ones that yielded the slow reaction times—than it did for the other two. When the attention check was removed or when its timing was equalized across conditions, reaction time effects

were eliminated, suggesting that the original pattern of findings may have been due to the confound.

If the standard for replication is significance of particular statistical tests at  $p < 0.05$ , then this experiment replicated successfully. But the effect estimates were inconsistent with the proposed theoretical explanation. A finding can be replicable without providing support for the underlying theory!

There's an important caveat to this story. The followup work *only* revealed that there was a confound in one particular experimental operationalization and did not provide evidence against automatic theory of mind in general. Indeed, others have suggested that different versions of this paradigm *do* reveal evidence for theory of mind processing once the confound is eliminated (El Kaddouri et al. 2020).

4150

## 4151 9.1 Experimental designs

4152 Experimental designs are fundamental to many fields; unfortunately,  
4153 the terminology used to describe them can vary, which can get quite  
4154 confusing! Here we will mostly describe an experiment as a relationship  
4155 between some manipulation(s), in which participants are randomly as-  
4156 signed to experimental conditions to estimate effects on some measure.

4157 Factors are the dimensions along which manipulations vary. For exam-  
4158 ple, in our case study above, the two factors were participant belief and

4159 agent belief. Another terminology it's good to be familiar with is the  
4160 terms used in chapters 5–7, which are often used in econometrics and  
4161 statistics: **treatment** (manipulation) and **outcome** (measure).<sup>2</sup>

4162 In this section, we'll discuss key dimensions on which experiments vary:  
4163 (1) how many factors they incorporate and how these factors are crossed;  
4164 (2) how many conditions and measures are given to each participant; and  
4165 (3) if manipulations have discrete levels or fall on a continuous scale.

### 4166 9.1.1 *A two-factor experiment*

4167 The classical “design of experiments” framework has as its goal to sep-  
4168 arate observed variability in the dependent measure into (1) variability  
4169 due to the manipulation(s) and (2) other variability, including measure-  
4170 ment error and participant-level variation. This framework maps nicely  
4171 onto the statistical framework described in chapters 5–7. In essence, this  
4172 framework models the distribution of the measure using the condition  
4173 structure of our experiment as the predictor.

4174 Different experimental designs will allow us to estimate specific effects  
4175 more and less effectively. Recall in chapter 5, we estimated the effect of  
4176 our tea/milk order manipulation by a simple subtraction:  $\beta = \theta_T - \theta_C$   
4177 (where  $\beta$  is the effect estimate, and  $\theta$ s indicate the estimates for each  
4178 condition, treatment  $T$  and control  $C$ ; we called them  $\theta_T$  and  $\theta_M$  in

<sup>2</sup> Terminology here is hard. In psychology people sometimes say there's an **independent variable** (the manipulation, which is causally prior and hence “independent” of other causal influences) and a **dependent variable** (the measure, which causally depends on the manipulation, or so we hypothesize). We find this terminology to be hard to remember because the terms are so different from the actual concepts being described.

that chapter to denote tea- and milk-first conditions). This logic works just fine also if there are two distinct treatments in a three-condition experiment: each treatment can be compared to control separately. For treatment 1,  $\beta_{T_1} = \theta_{T_1} - \theta_C$  and  $\beta_{T_2} = \theta_{T_2} - \theta_C$ .

This logic is going to get more complicated if we have more than one distinct factor of interest, though. Let's look at an example.

Young et al. (2007) were interested in how moral judgments depend on both the beliefs of actors and the outcomes of their actions. They presented participants with vignettes in which they learned, for example, that Grace visits a chemical factory with her friend and goes to the coffee break room, where she sees a white powder that she puts in her friend's coffee. They then manipulated both Grace's *beliefs* and the *outcomes* of her action following the schema in figure 9.3. Participants ( $N = 10$ ) used a four-point Likert scale to rate whether the actions were morally forbidden (1) or permissible (4).

		Outcome	
		Negative	Neutral
Belief	Negative	Grace thinks the powder is <b>toxic</b> . It is <b>toxic</b> . Her friend <b>dies</b> .	Grace thinks the powder is <b>toxic</b> . It is <b>sugar</b> . Her friend is <b>fine</b> .
	Neutral	Grace thinks the powder is <b>sugar</b> . It is <b>toxic</b> . Her friend <b>dies</b> .	Grace thinks the powder is <b>sugar</b> . It is <b>sugar</b> . Her friend is <b>fine</b> .

Figure 9.3  
The 2 x 2 crossed design used in Young et al. (2007)

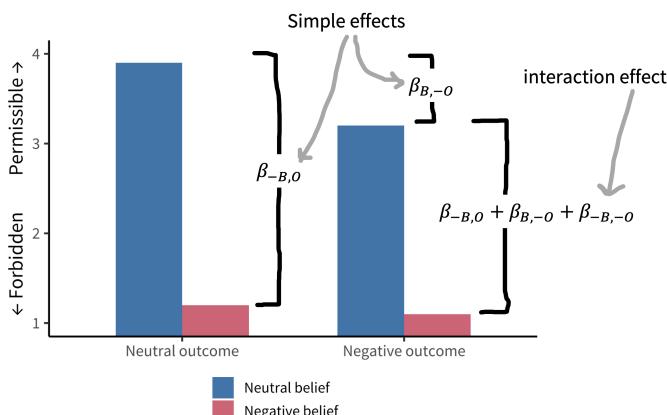


Figure 9.4  
Moral permissibility as a function of belief and outcome. Results from Young et al. (2007), annotated with the estimated effects. Simple effects measure differences between the individual conditions and the neutral belief, neutral outcome condition. The interaction measures the difference between the predicted sum of the two simple effects and the actual observed data for the negative belief, negative outcome condition.

4194 Young et al.'s design has two factors—belief and outcome—each with  
 4195 two levels (neutral and negative, noted as  $B$  and  $-B$  for belief and  $O$   
 4196 and  $-O$  for outcome).<sup>3</sup> These factors are **fully crossed**: each level of  
 4197 each factor is combined with each level of each other.

4198 This fully-crossed design makes it easy for us to estimate quantities of  
 4199 interest. Let's say that our **reference group** (equivalent to the control  
 4200 group for now) is neutral belief, neutral outcome. Now it's easy to use  
 4201 the same kind of subtraction we did before to estimate particular effects  
 4202 we care about. For example, we can look at the effect of negative belief  
 4203 in the case of a neutral outcome:  $\beta_{-B,O} = \theta_{-B,O} - \theta_{B,O}$ . This effect is  
 4204 shown on the left side of figure 9.4.

4205 But now there is a complexity: these two **simple effects** (effects of one  
 4206 variable at a particular level of another variable) together suggest that the  
 4207 combined effect  $\beta_{-B,-O}$  in the negative belief, negative outcome con-  
 4208 dition should be equal to the sum of  $\beta_{-B,O}$  and  $\beta_{B,-O}$ .<sup>4</sup> As we can see  
 4209 from figure 9.4, that's not right. If it were, the negative belief, negative  
 4210 outcome condition would be below the minimum possible rating!

4211 Instead, we observe an **interaction effect** (sometimes called a **two-way**  
 4212 **interaction** when there are two factors): the effect when both factors  
 4213 are present is different than the sum of the two simple effects. To cap-  
 4214 ture this effect, we need an interaction term:  $\beta_{-B,-O}$ .<sup>5</sup> In other words,

<sup>3</sup> Neither of these is necessarily a “control” condition: the goal is simply to compare these two levels of the factor—negative and neutral—to estimate the effect due to the factor.

<sup>4</sup> If you're interested, you can also compute the **average** or **main effect** of a particular factor via the same subtractive logic. For example, the average effect of negative belief ( $-B$ ) vs a neutral belief ( $B$ ) can be computed as  $\beta_{-B} = \frac{(\theta_{-O,-B} + \theta_{O,-B}) - (\theta_{-O,B} + \theta_{O,B})}{2}$ .

<sup>5</sup> If you're reading carefully, you might be thinking that this all sounds like we're talking about the analysis of variance (ANOVA), not about experimental design per se. These two topics are actually the same topic! The question is how to design an experiment so that these statistical models can be used to estimate particular effects—and combinations of effects—that we care about. In case you missed it, we discuss modeling interactions in a regression framework in chapter 7.

4215 the effect of negative beliefs (intent) on subjective moral permissibil-  
 4216 ity depends on whether the action caused harm. Critically, without a  
 4217 fully-crossed design, we can't estimate this interaction and we would  
 4218 have made an incorrect prediction about one condition.

4219 *9.1.2 Generalized factorial designs*

4220 Young et al.'s design, in which there are two factors with two levels  
 4221 each, is called a **2 x 2 design** (pronounced "two by two"). These 2 x 2  
 4222 designs are incredibly common and useful, but they are only one of an  
 4223 infinite variety of such designs that can be constructed.

4224 Say we added a third factor to Young et al.'s design such that Grace ei-  
 4225 ther feels neutral toward her friend or is angry on that day. If we fully  
 4226 crossed this third affective factor with the other two (belief and out-  
 4227 come), we'd have a 2 x 2 x 2 design. This design would have eight con-  
 4228 ditions:  $(A, B, O)$ ,  $(A, B, -O)$ ,  $(A, -B, O)$ ,  $(A, -B, -O)$ ,  $(-A, B, O)$ ,  
 4229  $(-A, B, -O)$ ,  $(-A, -B, O)$ ,  $(-A, -B, -O)$ . These conditions would  
 4230 in turn allow us to estimate both two-way and three-way interactions,  
 4231 enumerated in table 9.1.

Table 9.1  
 Effects in a 2 x 2 x 2 design with affect, belief, and outcome as factors.

Effect	Term Type
Affect	Main effect

Effect	Term Type
Belief	Main effect
Outcome	Main effect
Affect X Belief	2-way interaction
Affect X Outcome	2-way interaction
Belief X Outcome	2-way interaction
Affect X Belief X Outcome	3-way interaction

<sup>4232</sup> Three-way interactions are hard to think about! The affect X belief X

<sup>4233</sup> outcome interaction tells you about the difference in moral permissibil-

<sup>4234</sup> ity that's due to all three factors being present as opposed to what you'd

<sup>4235</sup> predict on the basis of your estimates of the two-way interactions. In

<sup>4236</sup> addition to being hard to think about, higher-order interactions tend to

<sup>4237</sup> be hard to estimate, because estimating them accurately requires you to

<sup>4238</sup> have a stable estimate of all of the lower-order interactions ([McClelland](#)

<sup>4239</sup> and [Judd 1993](#)). For this reason, we recommend against experimental

<sup>4240</sup> designs that rely on higher-order interactions unless you are in a situ-

<sup>4241</sup> ation where you both have strong predictions about these interactions

<sup>4242</sup> and are confident in your ability to estimate them appropriately.

<sup>4243</sup> Things can get even more complicated. If you have three factors with

<sup>4244</sup> two levels each, as in the example above (table 9.1), you can estimate

<sup>4245</sup> seven total effects of interest. But if you have *four* factors with two

4246 levels each, you get 15. Four factors with *three* levels each gets you a  
4247 horrifying 80 different effects!<sup>6</sup> This way lies madness, at least from the  
4248 perspective of estimating and interpreting individual effects in a reason-  
4249 able sample size. Again, we suggest starting with one- and two-factor  
4250 designs. There is a lot to be learned from simple designs that follow  
4251 good measurement and sampling practices.

<sup>6</sup> The general formula for  $N$  factors with  $M$  levels each is  $M^N - 1$ .

#### DEPTH

### *Estimation strategies for generalized factorial designs*

So, what should you do if you really do care about four or more factors—in the sense that you want to estimate their effects and include them in your theory? The simplest strategy is to start your research off by measuring them independently in a series of single-factor experiments. This kind of setup is natural when there is a single reference level for each factor of interest, and such experiments can provide a basis for judging which factors are most important for your outcome and, hence, which should be prioritized for experiments to estimate interactions.

On the other hand, sometimes there is no reference level for a factor. For example, in the Kovács, Téglás, and Endress (2010) paradigm, it's not clear whether a positive or negative belief is the reference level. That's not a problem in a fully-crossed design like theirs, but this situation can pose a problem if you have more than two such factors. Ideally you would want to run independent experiments, but you have to choose some level for all

of the other variables—you can't just assume that one level is “neutral.”

One solution that lets you compute main effects but not interactions is called a **Latin square**. Latin squares are a good solution for three-factor designs, which is the level at which a fully-crossed design typically gets overwhelming. A Latin square is an  $n \times n$  matrix in which each number occurs exactly once in each row and column, for example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix}$$

This Latin square for  $n = 3$  gives the solution for how to balance factors across a  $3 \times 3 \times 3$  experiment. The row number is one factor, the column number is the second factor, and the number in the cell is the third factor. So one condition would be (1,1,1), the first level of all factors, shown in the upper left cell. Another would be (3,3,2), the lower right cell. Although a fully-crossed design would require 27 cells to be run, the Latin square has only nine. Critically, the combinations of factors are balanced across the nine cells so that the average effect of each level of the three factors can be estimated.

There are also fancier methods available. For example, the literature on **optimal experiment design** contains methods for choosing the most informative sequence of experiments to run in order to estimate the parameters in a model that can include many factors and their interactions (Myung and Pitt 2009). Going down this road typically means having

an implemented computational theory of your domain, but it can be a very productive strategy for exploring a complex experimental space with many factors.

4254

### 4255 9.1.1 Between- vs within-participant designs

4256 Once you know what factor(s) you would like to manipulate in your  
 4257 experiment, the next step is to consider how these will be presented  
 4258 to participants, and how that presentation will interact with your mea-  
 4259 surements. The biggest decision to be made is whether each participant  
 4260 will experience one level of a factor—a **between-participants design**—or  
 4261 whether they will experience multiple levels—a **within-participants de-  
 4262 sign**. Figure 9.5 shows a simple example of between-participants design  
 4263 with four participants (two assigned to each condition), while figure 9.6  
 4264 shows a within-participants version of the same design.

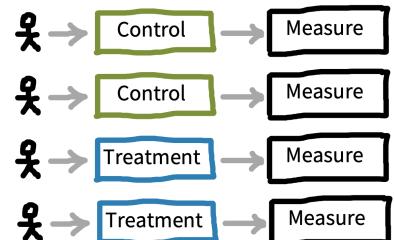


Figure 9.5  
 A between-participants design.

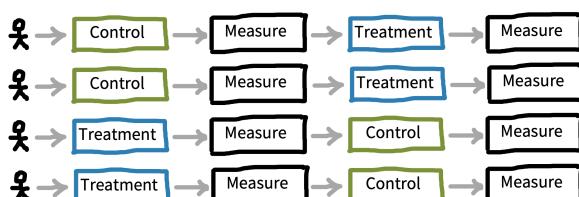


Figure 9.6  
 A within-participants design, counter-  
 balanced for order (discussion of coun-  
 terbalancing below).

4265 Because people are very variable, the decision whether to measure a par-  
 4266 ticular factor between- or within-participants is consequential. Imag-  
 4267 ine we’re estimating our treatment effect as before, simply by comput-  
 4268 ing  $\hat{\beta} = \hat{\theta}_T - \hat{\theta}_C$  with each of these estimates from different populations

4269 of participants. In this scenario, our estimate  $\hat{\beta}$  contains three compo-  
4270 nents: (1) the true differences between  $\theta_T$  and  $\theta_C$ , (2) sampling-related  
4271 variation in which participants from the population ended up in the sam-  
4272 ples for the two conditions, and (3) measurement error. Component 2  
4273 is present because any two samples of participants from a population  
4274 will differ in their average on a measure—this is precisely the kind of  
4275 sampling variation we saw in the null distributions in chapter 6.

4276 When our experimental design is within participants, component 2 is  
4277 not present because participants in both conditions are sampled from the  
4278 *same* population. If we get unlucky and all of our participants are lower  
4279 than the population mean on our measure, then that unluckiness affects  
4280 our conditions equally. The consequences for choosing an appropriate  
4281 sample size are fairly extreme: between-participants designs typically  
4282 require between two and eight times as many participants as within-  
4283 participants designs!<sup>7</sup>

4284 Given these advantages, why would you consider using a between-  
4285 participants design? A within-participants design is simply not possible  
4286 for all experiments. For example, consider a medical intervention—say,  
4287 a new surgical procedure that is being compared to an established  
4288 one. Patients cannot receive two different procedures, and so no  
4289 within-participant comparison is possible.

<sup>7</sup> If you want to estimate how big an advantage you get from within-participants data collection, you need to know how correlated (reliable) your observations are. One analysis of this issue (Lakens 2016) suggests that the key relationship is that  $N_{within} = N_{between}(1 - \rho)/2$  where  $\rho$  is the correlation between the measurement of the two conditions within individuals. The more correlated they are, the smaller your within-participants  $N$ .

4290 Most manipulations in the behavioral sciences are not so extreme, but  
4291 it still may be impractical or inadvisable to deliver multiple conditions.

4292 Greenwald (1976) distinguishes three types of undesirable effects:<sup>8</sup>

4293 – **Practice effects** occur when administering the measure or the  
4294 treatment will lead to change. Imagine a curriculum intervention  
4295 for teaching a math concept—it would be hard to convince a  
4296 school to teach the same topic to students twice, and the effect  
4297 of the second round of teaching would likely be quite different  
4298 than the first!

4299 – **Sensitization effects** occur when seeing two versions of an  
4300 intervention mean that you might respond differently to the  
4301 second than the first because you have compared them and  
4302 noticed the contrast. Consider a study on room lighting—if the  
4303 experimenters are constantly changing the lighting, participants  
4304 may become aware that lighting is the focus of the study!

4305 – **Carryover effects** refer to the case where one treatment might  
4306 have a longer-lasting effect than the measurement period. For  
4307 example, imagine a study in which one treatment was to make  
4308 participants frustrated with an impossible puzzle; if a second con-  
4309 dition were given after this first one, participants might still be  
4310 frustrated, leading to spillover of effects between conditions.

<sup>8</sup> We tend to think of all of these as being forms of carryover effect, and sometimes use this label as a catch-all description. Some people also use the picturesque description “poisoning the well” (Gelman 2017)—earlier conditions “ruin” the data for later conditions.

4311 All of these issues can lead to real concerns with respect to within-  
4312 participant designs. But the desire for effect estimates that are  
4313 completely unbiased by these concerns may lead to the overuse of  
4314 between-participant designs (Gelman 2017). As we mentioned above,  
4315 between-participant designs come at a major cost in terms of power  
4316 and precision.

4317 An alternative approach is to acknowledge the possibility of carryover  
4318 type effects and seek to mitigate them. First, you can make sure that the  
4319 order of condition is randomized or balanced (see below); and second,  
4320 you can analyze these carryover effects within your statistical model (for  
4321 example by estimating the interaction of condition and order).<sup>9</sup>

4322 We summarize the state of affairs from our perspective in figure 9.7. We  
4323 think that within-participant designs should be preferred whenever pos-  
4324 sible. This conclusion is also consistent with meta-research we've done  
4325 on replications from our course: across 176 student replications, the use  
4326 of a within-subjects design was the strongest correlate of a successful  
4327 replication (Boyce, Mathur, and Frank 2023).<sup>10</sup>

4328 *9.1.2 Repeated measures and experimental items*

4329 We just discussed decision-making about whether to administer mul-  
4330 tiple *manipulations* to a single participant. An exactly analogous deci-

<sup>9</sup> Even when one factor must be varied between participants, it is often still possible to vary others within subjects, leading to a **mixed design** in which some factors are between and others within.

<sup>10</sup> Caveat: this study used an observational design, so no causal inference is possible.

- | <b>Between</b>  | <b>Within</b>  |
|---|--|
| <ul style="list-style-type: none"> <li>• Main advantage           <ul style="list-style-type: none"> <li>• No contamination by other exposure to experimental materials</li> </ul> </li> <li>• Disadvantages           <ul style="list-style-type: none"> <li>• Requires many participants</li> <li>• Individual differences create a lot of variability in groups</li> <li>• Potential for assignment bias: need to control for differences between groups</li> <li>• Other environmental group differences</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>• Main advantage           <ul style="list-style-type: none"> <li>• Eliminates subject variability</li> <li>• Relatively few participants needed, because of this lack of variability</li> </ul> </li> <li>• Disadvantages           <ul style="list-style-type: none"> <li>• Carryover effects mean that ordering of conditions can be problematic</li> <li>• Not always possible</li> </ul> </li> <li>• General contention: preferable when possible</li> </ul> |

Figure 9.7

Pros and cons of between- vs within-participant designs. We recommend within-participant designs when possible.

4331 sion comes up for *measures*! And our take-home will be similar: unless  
 4332 there are specific difficulties that come up, it's usually a very good idea  
 4333 to make multiple measurements (via multiple experimental trials) for  
 4334 each participant in each condition.

4335 You can create a between-participants design where you administer  
 4336 your manipulation and then measure multiple times. This scenario is  
 4337 pictured in figure 9.8). Sometimes this works quite well. For example,  
 4338 imagine a transcranial magnetic stimulation (TMS) experiment: partic-  
 4339 ipants receive neural stimulation for a period of time, targeted at a par-  
 4340 ticular region. Then they perform some measurement task repeatedly  
 4341 until it wears off. The more times they perform the measurement task,  
 4342 the better the estimate of whatever effect (when compared to a control  
 4343 of TMS to another region, say).

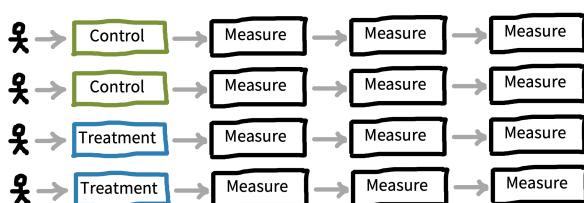


Figure 9.8

A between-participants, repeated-measures design.

4344 Sometimes this design is called a **repeated measures** design, but termino-  
4345 nology here is tricky again. The term “repeated measures” refers to any  
4346 experiment where each participant is measured more than once, includ-  
4347 ing both between-participants *and* within-participants designs.<sup>11</sup> Our  
4348 advice is *both* to use within-participants designs *and* to get multiple mea-  
4349 surements from each participant.

4350 Why? In the last subsection, we described how variability in our esti-  
4351 mates in a between-participants design depends on three components:  
4352 (1) true condition differences; (2) sampling variation between condi-  
4353 tions; and (3) measurement error.

4354 Within-participants designs are good because they don’t include (2).  
4355 Repeated measures reduce (3): the more times you measure, the lower  
4356 your measurement error, leading to greater measure reliability!

4357 There are problems with repeating the same measure many times, how-  
4358 ever. Some measures can’t be repeated without altering the response.

4359 To take an obvious example, we can’t give the exact same math prob-  
4360 lem twice and get two useful measurements of mathematical ability!

4361 The typical solution to this problem is to create multiple items. In the  
4362 case of a math assessment, you create multiple problems that you believe  
4363 test the same concept but have different numbers or other superficial  
4364 characteristics.

<sup>11</sup> We’re talking about multiple trials with the same measure, not multiple distinct measures. As we discussed in chapter 8, we tend to be against measuring lots of different things in a single experiment—in part because of the concerns that we’re articulating in this chapter: if you have time, it’s better to make more precise measures of what you care about most. Measuring one thing well is hard enough. Much better to measure one thing well than many things badly.

4365 Using multiple items for measurement is good for two reasons. First, it  
 4366 reduces measurement error by allowing responses to be combined across  
 4367 items. But second, it increases the generalizability of the measurement.  
 4368 An effect that is consistent across many different items is more likely  
 4369 to be an effect that can be generalized to a whole class of stimuli—in  
 4370 precisely the same way that the use of multiple participants can license  
 4371 generalizations across a population of people (Clark 1973).

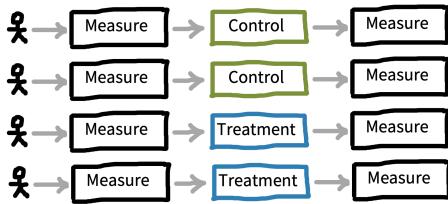


Figure 9.9  
 A between-participants, pre-post design.

4372 One variation on the repeated measures, between-participants design is  
 4373 a specific version where the measure is administered both before (pre-)  
 4374 and after (post-) intervention, as in figure 9.9. This design is sometimes  
 4375 known as a **pre-post** design. It is extremely common in cases where  
 4376 the intervention is larger scale and harder to give within-participants,  
 4377 such as in a field experiment where a policy or curriculum is given to  
 4378 one sample and not to another. The premeasurements can be used to  
 4379 subtract out participant-level variability and recover a more precise es-  
 4380 timate of the treatment effect. Recall that our treatment effect in a pure  
 4381 between-participants design is  $\beta = \theta_T - \theta_C$ . In a pre-post design, we  
 4382 can do better by computing  $\beta = (\theta_{T_{post}} - \theta_{T_{pre}}) - (\theta_{C_{post}} - \theta_{C_{pre}})$ .  
 4383 This equation says, “How much more did the treatment group go up

4384 than the control group?<sup>12</sup>

4385 In sum, within-participants, repeated-measurement designs are the  
4386 bread and butter of most research in perception, psychophysics, and  
4387 cognitive psychology. When both manipulations and measures can be  
4388 repeated, these designs afford high measurement precision even with  
4389 small sample sizes; they are recommended whenever possible.

<sup>12</sup> This estimate is sometimes called a “difference in differences.” The basic idea is widely used in the field of econometrics, both in experimental and quasi-experimental cases (Cunningham 2021). In practice, though, we recommend using the pre-treatment measurements as a covariate in a model-based analysis, not just doing the simple subtraction.

### ⚠ ACCIDENT REPORT

#### *Stimulus-specific effects*

Imagine you’re a psycholinguist who has the hypothesis that nouns are processed faster than verbs. You run an experiment where you pick out ten verbs and ten nouns, then measure a large sample of participants’ reading time for each of these. You find strong evidence for the predicted effect and publish a paper on your claim. The only problem is that, at the same time, someone else has done exactly the same study—with different nouns and verbs—and published a paper making the opposite claim. When this happens, it is possible that each effect is driven by the specific experimental items that were chosen, rather than a generalization that is true of nouns and verbs in general (Clark 1973).

The problem of generalization from sample to population is not new—as we discussed in chapter 6, we are constantly making this kind of inference with the samples of people that participate in our experiments. Our classic statistical techniques are designed to quantify our ability to generalize

from a sample of participants to a population, so we recognize that a very small sample size leads to a weak generalization. The exact same issue comes up with *items*: a very small sample of experimental items leads to a weak generalization to the population of items.

Item effects are kind of like accidentally finding a group of ten people whose left toes are longer than their right ones. If you continued to measure the same group's toes, you could continue to replicate the difference in length. But that doesn't mean it's true of the population as a whole.

This kind of **stimulus generalizability** problem comes up across many different areas of psychology. In one example, hundreds of papers were written about a phenomenon called the “risky shift”—in which groups deliberating about a decision would produce riskier decisions than individuals. Unfortunately, this phenomenon appeared to be completely driven by the specific choice of vignettes that groups deliberated about, with some stories producing a risky shift and others producing a more conservative shift (Westfall, Judd, and Kenny 2015).

Another example comes from the memory literature, where in a classic paper, Baddeley, Thomson, and Buchanan (1975) suggested that words that take longer to pronounce (“tycoon” or “morphine”) would be remembered worse than words that took a shorter amount of time (“ember” or “wicket”) even when they had the same number of syllables. This effect also appears to be driven by the specific sets of words chosen in the original paper. It’s very replicable with that particular stimulus set but

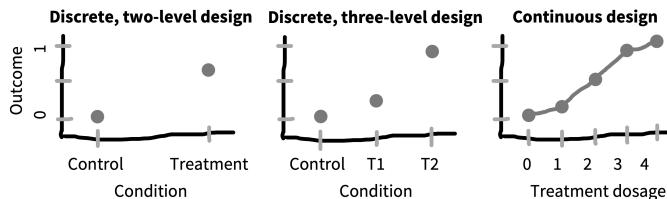
not generalizable across other sets (Lovatt, Avons, and Masterson 2000).

The implication of these examples is clear: experimenters need to take care in both their experimental design and analysis to avoid overgeneralizing from their stimuli to a broader construct. Three primary steps can help experimenters avoid this pitfall:

1. To maximize generality, use samples of experimental items—words, pictures, or vignettes—that are comparable in size to your samples of participants.
2. When replicating an experiment, consider taking a new sample of items as well as a new sample of participants. It's more work to draft new items, but it will lead to more robust conclusions.
3. When experimental items are sampled at random from a broader population, use a statistical model that includes this sampling process (e.g., mixed effects models with random intercepts for items from chapter 7).

4393 9.1.1 Discrete and continuous experimental manipulations

4394 Most experimental designs in psychology use discrete condition manipulations: treatment vs control. In our view, this decision often leads  
4395 to a lost opportunity relative to a more continuous manipulation of the  
4396 strength of the treatment. The goal of an experiment is to estimate a  
4397 causal effect; ideally, this estimate can be generalized to other contexts  
4398 and used as a basis for theory. Measuring not just one effect but instead  
4399 a **dose–response** relationship—how the measure changes as the strength  
4400 of the manipulation is changed—has a number of benefits in helping to  
4401 achieve this goal.



4402 Many manipulations can be **titrated**—that is, their strength can be varied  
4403 continuously—with a little creativity on the part of an experimenter. A  
4404 curriculum intervention can be applied at different levels of intensity,  
4405 perhaps by changing the number of sessions in which it is taught. For a  
4406 priming manipulation, the frequency or duration of prime stimuli can  
4407 be varied. Two stimuli can be morphed continuously so that categorization boundaries can be examined.<sup>13</sup>

4410 Dose–response designs are useful because they provide insight into

Figure 9.10

Three schematic designs. (left) Control and treatment are two levels of a nominal variable. (middle) Control is compared to ordered levels of a treatment. (right) Treatment level is an interval or ratio variable such that points can be connected and a parametric curve can be extrapolated.

<sup>13</sup> These methods are extremely common in perception and psychophysics research, in part because the dimensions being studied are often continuous in nature. It would be basically impossible to estimate a participant's visual contrast sensitivity *without* continuously manipulating the contrast of the stimulus!

the shape of the function mapping your manipulation to your measure. Knowing this shape can inform your theoretical understanding!

Consider the examples given in figure 9.10. If you only have two conditions in your experiment, then the most you can say about the relationship between your manipulation and your measure is that it produces an effect of a particular magnitude; in essence, you are assuming that condition is a nominal variable. If you have multiple ordered levels of treatment, you can start to speculate about the nature of the relationship between treatment and effect magnitude. But if you can measure the strength of your treatment, then you can begin to describe the nature of the relationship between the strength of treatment and strength of effect via a parametric function (e.g., a linear regression, a sigmoid, or other function).<sup>14</sup> These parametric functions can in turn allow you to generalize from your experiment, making predictions about what would happen under intervention conditions that you didn't measure directly!

#### DEPTH

##### *Trade-offs associated with titrated designs*

Like adults, babies like to look at more interesting, complex stimuli. But do they uniformly prefer complex stimuli, or do they search for stimuli at an appropriate level of complexity for their processing abilities? To

<sup>14</sup> These assumptions are theory-laden, of course—the choice of a linear function or a sigmoid is not necessary: nothing guarantees that simple, smooth, or monotonic functions are the right ones. The important point is that choosing a function makes explicit your assumptions about the nature of the treatment–effect relationship.

test this hypothesis, Brennan, Ames, and Moore (1966) exposed infants in three different age groups (3, 8, and 14 weeks,  $N = 30$ ) to black and white checkerboard stimuli with three different levels of complexity (2 x 2, 8 x 8, and 24 x 24).

Their findings are plotted in figure 9.11: the youngest infants preferred the simplest stimuli, while infants at an intermediate age preferred stimuli of intermediate complexity, and the oldest infants preferred the most complex stimuli. These findings help to motivate the theory that infants attend preferentially to stimuli that provide appropriate learning input for their processing ability (Kidd, Piantadosi, and Aslin 2012).

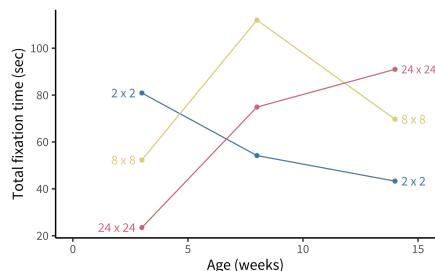


Figure 9.11  
Infants' looking time, plotted by stimulus complexity and infant age. Data from Brennan, Ames, and Moore (1966).

If your goal is simply to detect whether an effect is zero or nonzero, then dose-response designs do not achieve the maximum statistical power. For example, if Brennan, Ames, and Moore (1966) simply wanted to achieve maximal statistical power, they probably should have only tested two age groups and two levels of complexity (say, 3 and 14 week infants and 2 x 2 and 24 x 24 checkerboards). That would have been enough to show an interaction of complexity and age, and their greater resources devoted to

these four (as opposed to nine) conditions would mean more precise estimates of each. But their findings would be less clearly supportive of the view that infants prefer stimuli that are appropriate to their processing ability, because no group would have preferred an intermediate level of complexity (as the nine-week-olds apparently did). By seeking to measure intermediate conditions, they provided a stronger test of their theory.

4429

## 4430 9.2 Choosing your manipulation

4431 In the previous section, we reviewed a host of common experimental  
4432 designs. These designs provide a palette of common options for combin-  
4433 ing manipulations and measures. But your choice must be predicated on  
4434 the specific manipulation you are interested in! In this section, we dis-  
4435 cuss considerations for experimenters as they design manipulations.

4436 In chapter 8, we talked about *measurement* validity, but the idea of va-  
4437 lidity concept can be applied to manipulations as well as measures. In  
4438 particular, a manipulation is valid if it corresponds to the construct that  
4439 the experimenter intends to intervene on. In this context, *internal* va-  
4440 lidity threats to manipulations tend to refer to cases where factors in  
4441 the experimental design keep the intended manipulation from actually  
4442 intervening on the construct of interest. In contrast, *external* validity

4443 threats to manipulations tend to be cases where the manipulation sim-  
4444 ply doesn't line up well with the construct of interest.

4445 *9.2.1 Internal validity threats: Confounding*

4446 First and foremost, manipulations must actually manipulate the con-  
4447 struct whose causal effect is being estimated. If they *actually* manipulate  
4448 something else instead, they are **confounded**. This term is used widely  
4449 in psychology, but it's worth revisiting what it means. An **experimental**  
4450 **confound** is a variable that is created in the course of the experimental  
4451 design that is both causally related to the predictor and potentially also  
4452 related to the outcome. As such, it is a threat to **internal validity**.

4453 Let's go back to our discussion of causal inference in chapter 1. Our  
4454 goal was to use a randomized experiment to estimate the causal effect of  
4455 money on happiness. But just giving people money is a big intervention  
4456 that involves contact with researchers—contact alone can lead to an ex-  
4457 perimental effect even if your manipulation fails. For that reason, many  
4458 studies that provide money to participants either give a small amount  
4459 of money or a large amount of money. This design keeps researcher  
4460 contact consistent in both conditions, implying that the difference in  
4461 outcomes between these two conditions should be due to the amount  
4462 of money received (unless there are other confounds!).

4463 Suppose you were designing an experiment of this sort and you wanted  
 4464 to follow our advice and use a within-participants design. You could  
 4465 measure happiness, give participants \$100, wait a month and measure  
 4466 happiness again, give participants \$1,000, wait a month, and then mea-  
 4467 sure happiness for the third time. The trouble is, this design has an  
 4468 obvious experimental confound (figure 9.12): the order of the mone-  
 4469 tary gifts. Maybe happiness just went up more over time, irrespective  
 4470 of getting the second gift.

4471 If you think your experimental design might have a confound, you  
 4472 should think about ways to remove it. A first option is **elimination**,  
 4473 which we described above: basically, matching a particular variable  
 4474 across different conditions. This should be our first option for most con-  
 4475 founds. Unfortunately, in our within-participants money-happiness  
 4476 study, order is confounded with condition so if we match orders we  
 4477 have eliminated our condition manipulation entirely.

4478 A second option is **counterbalancing**, in which we vary a confounding  
 4479 factor systematically across participants so its average effect is zero across  
 4480 the whole experiment. In the case of our example, counterbalancing  
 4481 order across participants is a very safe choice. Some participants get  
 4482 \$100 first and others get \$1,000 first. That way, you are guaranteed that  
 4483 the order of conditions will have no effect of the confound on your

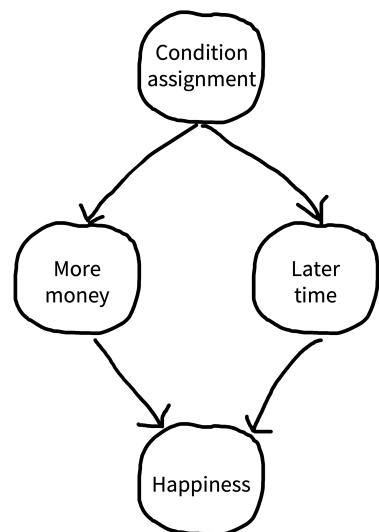


Figure 9.12  
 Confounding order and condition as-  
 signment means that you can't make an  
 inference about the link between money  
 and happiness.

average effect. The effect of this counterbalancing is that it “snips” the causal dependency between condition assignment and later time. We notate this on our causal diagram with a scissors icon (figure 9.13).<sup>15</sup> Time can still have an effect on happiness, but the effect is independent from the effect of condition and, hence, your experiment can still yield an unbiased estimate of the condition effect.

Counterbalancing gets trickier when you have too many levels on a variable or multiple confounding variables. In that case, it may not be possible to do a full counterbalance so that all combinations of these factors are seen by equal numbers of participants. You may have to rely on partial counterbalancing schemes or Latin square designs (see the DEPTH box above; in this case, the Latin squares are used to create orderings of stimuli such that the position of each treatment in the order is controlled across two other confounding variables).

A final option, especially useful for such tricky cases, is **randomization**—that is, choosing which level of a nuisance variable to administer to the participant via a random choice. Randomization is increasingly common now that many experimental interventions are delivered by software. If you *can* randomize experimental confounds, you probably should. The only time you really get in trouble with randomization is when you have a large number of options, a small number of partic-

<sup>15</sup> In practice, counterbalancing is like adding an additional factor to your factorial design! But because the factor is a **nuisance factor**—basically, one we don’t care about—we don’t discuss it as a true condition manipulation. Despite that, it’s a good practice to check for effects of these sorts of nuisance factors in your preliminary analysis. Even though your average effect won’t be biased by it, it introduces variation that you might want to understand to interpret other effects and plan new studies.

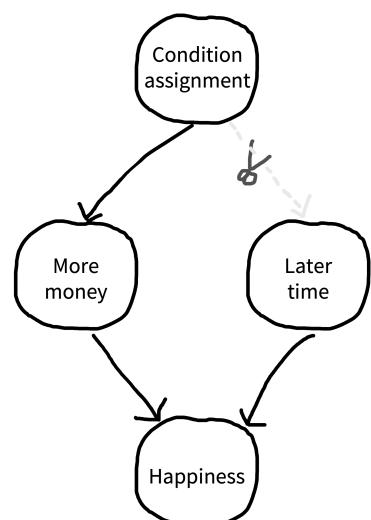


Figure 9.13  
Confounding between a specific condition and the time at which it's administered can be removed by counterbalancing or randomization of order.

4505 ipants, or some combination of the two. Then you can end up with  
4506 unbalanced levels of the randomized factors. Averaging across many ex-  
4507 periments, a lack of balance will come out in the wash, but in a single  
4508 experiment, it can lead to unfortunate bias in numbers.

4509 A good approach to thinking through your experimental design is to  
4510 walk through the experiment step by step and think about potential  
4511 confounds. For each of these confounds, consider how it might be  
4512 removed via counterbalancing or randomization. As our case study  
4513 shows, confounds are not always obvious, especially in complex  
4514 paradigms. There is no sure-fire way to ensure that you have spotted  
4515 every one—sometimes the best way to avoid them is simply to present  
4516 your candidate design to a skeptical friend.

#### 4517 9.2.2 Internal validity threats: Placebo, demand, and expectancy

4518 A second class of important threats to internal validity comes from cases  
4519 where the research design is confounded by factors related to how the  
4520 manipulation is administered, or even *that* a manipulation is adminis-  
4521 tered. In some cases, these create confounds that can be controlled; in  
4522 others they must simply be understood and guarded against. Rosnow  
4523 and Rosenthal (1997) called these “artifacts”: systematic errors related  
4524 to research *on* people, conducted *by* people.

4525 A placebo effect is a positive effect on the measure that comes as a re-  
4526 sult of participants' expectations about a treatment in the context of a  
4527 research study. The classic example of a placebo is medical: giving an  
4528 inactive sugar pill as a "treatment" leads some patients to report a reduc-  
4529 tion in whatever symptom they are being treated for. Placebo effects are  
4530 a major concern in medical research as well as a fixture in experimental  
4531 designs in medicine (Benedetti 2020). The key insight is that treatments  
4532 must not simply be compared to a baseline of no treatment but rather  
4533 to a baseline in which the psychological aspects of treatment are present  
4534 but the "active ingredient" is not. In the terms we have been using, the  
4535 experience of receiving a treatment (independent of the content of the  
4536 treatment) is a confounding factor when you simply compare treatment  
4537 to no treatment conditions.

#### ⚠️ ACCIDENT REPORT

### *Brain training?*

Can doing challenging cognitive tasks make you smarter? In the late 2000s and early 2010s, a large industry for "brain training" emerged. Companies like Lumos Labs, CogMed, BrainHQ, and CogniFit offered games, often modeled on cognitive psychology tasks, that claimed to lead to gains in memory, attention, and problem-solving.

These companies were basing their claims in part on a scientific literature reporting that concerted training on difficult cognitive tasks could lead to

benefits that transferred to other cognitive domains. Among the most influential of these was a study by Jaeggi et al. (2008). They conducted four experiments in which participants ( $N = 70$  across the studies) were assigned to either working memory training via a difficult working memory task (the “dual N-back”) or a no-training control, with training varying from eight days all the way to 19 days.

The finding from this study excited a tremendous amount of interest because they reported not only gains in performance on the specific training task but also on a general intelligence task that the participants had trained on. While the control group’s scores on these tasks improved, presumably just from being tested twice, there was a condition by time (pre-test vs post-test) interaction such that the scores of the trained groups (consolidated across all four training experiments) grew significantly more over the training period (figure 9.14). These results were interpreted as supporting transfer—whereby training on one task leads to broader gains—a key goal for “brain training.”

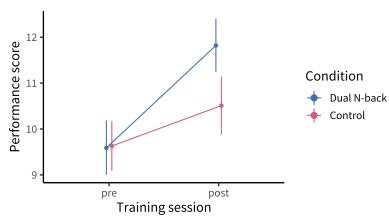


Figure 9.14  
The primary outcome graph for data from Jaeggi et al. (2008).

Careful readers of the original paper noticed signs of analytic flexibility (as discussed in chapters 3 and 6), however. For example, the key statistical model was fit to dataset created by post hoc consolidation of ex-

periments, which yielded  $p = 0.025$  on the key interaction (Redick et al. 2013). When data were disaggregated, it was clear that the measures and effects had differed in each of the different subexperiments (figure 9.15).

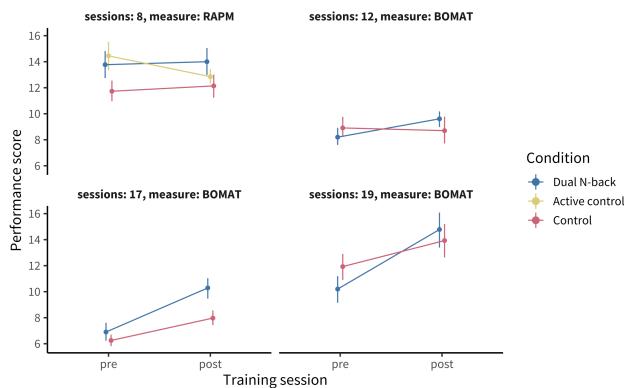


Figure 9.15  
The four subexperiments of Jaeggi et al. (2008), now disaggregated. Panels show 8-, 12-, 17-, and 19-session studies. Note the different measures: RAPM = Raven's Advanced Progressive Matrices; BOMAT = Bochumer Matrizentest. Based on Redick et al. (2013).

Several replications by the same group addressed some of these issues, but they still failed to show convincing evidence of transfer. In particular, there was no comparison to an **active control group** in which participants did some kind of alternative activity for the same amount of time (Simmons et al. 2016). Such a comparison is critical because a comparison to a **passive control group** (a group that does no intervention) confounds participants' general effort and involvement in the study with the specific training being used. Successful transfer compared to passive control could be the result of participants' involvement, expectations, or motivation rather than brain training per se.

A careful replication of the training study ( $N = 74$ ) with an active control

group and a wide range of outcome measures failed to find any transfer effects from working-memory training (Redick et al. 2013). A meta-analysis of 23 studies concluded that their findings cast doubt on working memory training for increasing cognitive functioning (Melby-Lervåg and Hulme 2013). In one convincing test of the cognitive transfer theory, a BBC show (“Bang Goes the Theory”) encouraged its listeners to participate in a six-week online brain-training study. More than 11,000 listeners completed the pre- and post-tests and at least two training sessions. Neither focused training of planning and reasoning nor broader training on memory, attention, and mathematics led to transfer to untrained tasks.

Placebo effects are one plausible explanation for some positive findings in the brain-training literature. Foroughi et al. (2016) recruited participants to participate via two different advertisements. The first advertised that “numerous studies have shown working memory training can increase fluid intelligence” (“placebo treatment” group), while the second simply offered experimental credits (control group). After a single training session, the placebo treatment group showed significant improvements to their matrix reasoning abilities. Participants in the placebo treatment group realized gains from training out of proportion with any they could have realized through training. Further, those participants who responded to the placebo treatment ad tended to endorse statements about the malleability of intelligence, suggesting that they might have been especially likely to self-select into the intervention.

Summarizing the voluminous literature on brain training, Simons et al.

(2016) wrote: “Despite marketing claims from brain-training companies of ‘proven benefits’ … we find the evidence of benefits from cognitive brain training to be ‘inadequate.’ ”

4542

4543 If placebo effects reflect what participants expect from a treatment, then

4544 **demand characteristics** reflect what participants think *experimenters*

4545 want and their desire to help the experimenters achieve that goal

4546 (Orne 1962). Demand characteristics are often raised as a reason for

4547 avoiding within-participants designs—if participants become alert to

4548 the presence of an intervention, they may then respond in a way that

4549 they believe is helpful to the experimenter. Typical tools for control-

4550 ling or identifying demand characteristics include using a cover story

4551 to mask the purpose of an experiment, using a debriefing procedure

4552 to probe whether participants typically guessed the purpose of an

4553 experiment, and (perhaps most effectively) creating a control condition

4554 with similar demand characteristics but missing a key component of

4555 the experimental intervention. Note that if you use a cover story

4556 to mask the purpose of your experiment, it’s worth thinking about

4557 whether you are using deception, which can raise ethical issues (see

4558 chapter 4). Certainly you should be sure to debrief participants about

4559 the true function of the experiment!

4560 The final entry into this list of internal validity threats is **experimenter**

4561 expectancy effects, where the experimenter's behavior biases partici-  
4562 pants in a way that results in the appearance of condition differences  
4563 where no true difference exists. The classic example of such effects is  
4564 from the animal learning literature and the story of Clever Hans. Clever  
4565 Hans was a horse who appeared able to do arithmetic by tapping out so-  
4566 lutions with his hoof. On deeper investigation, it became apparent he  
4567 was being cued by his trainer's posture (apparently without the trainer's  
4568 knowledge) to stop tapping when the desired answer was reached. The  
4569 horse knew nothing about math, but the experimenter's expectations  
4570 were altering the horse's behavior across different conditions.

4571 In any experiment delivered by human experimenters who know what  
4572 condition they are delivering, condition differences can result from ex-  
4573 perimenters imparting their expectations. Table 9.2 shows the results  
4574 of a meta-analysis estimating sizes of expectancy effects in a range of  
4575 domains—the magnitudes are shocking. There's no question that ex-  
4576 perimenter expectancy is sufficient to “create” many interesting phe-  
4577 nomena artifactually. The mechanisms of expectancy are an interesting  
4578 research topic in their own right; in many cases expectancies appear to  
4579 be communicated nonverbally in much the same way that Clever Hans  
4580 learned (Rosnow and Rosenthal 1997).

Table 9.2  
Magnitudes of expectancy effects. Based on R. Rosenthal (1994).

Domain	d	r	Example of type of study
Laboratory interviews	0.14	0.07	Effects of sensory restriction on reports of hallucinatory experiences
Reaction time	0.17	0.08	Latency of word associations to certain stimulus words
Learning and ability	0.54	0.26	IQ test scores, verbal conditioning (learning)
Person perception	0.55	0.27	Perception of other people's success
Inkblot tests	0.84	0.39	Ratio of animal to human Rorschach responses
Everyday situations	0.88	0.40	Symbol learning, athletic performance
Psychophysical judgments	1.05	0.46	Ability to discriminate tones
Animal learning	1.73	0.65	Learning in mazes and Skinner boxes
<i>Weighted mean</i>	0.70	0.33	
<i>Unweighted mean</i>	0.74	0.35	
<i>Median</i>	0.70	0.33	

<sup>4581</sup> In medical research, the gold standard is an experimental design where

4582 neither patients nor experimenters know which condition the patients  
4583 are in.<sup>16</sup> Results from other designs are treated with suspicion because  
4584 of their vulnerability to demand and expectancy effects. In psychology,  
4585 the most common modern protection against experimenter expectancy  
4586 is the delivery of interventions by a computer platform that can give  
4587 instructions in a coherent and uniform way across conditions.

4588 In the case of interventions that must be delivered by experimenters,  
4589 ideally experimenters should be unaware of which condition they are  
4590 delivering. On the other hand, the logistics of maintaining experi-  
4591 menter ignorance can be quite complicated in psychology. For this  
4592 reason, many researchers opt for lesser degrees of control: for example,  
4593 choosing to standardize delivery of an intervention via a script. These  
4594 designs are sometimes necessary for practical reasons but should be  
4595 scrutinized closely. “How can you rule out experimenter expectancy  
4596 effects?” is an uncomfortable question that should be asked more  
4597 frequently in seminars and paper reviews.

#### 4598 9.2.1 *External validity of manipulations*

4599 The goal of a specific experimental manipulation is to operationalize  
4600 a particular causal relationship of interest. Just as the relationship be-  
4601 tween measure and construct can be more or less valid, so too can the

<sup>16</sup> These are commonly referred to as double-blind designs (though the term masked is now often preferred).

4602 relationship between manipulation and construct. How can you tell?

4603 Just like in the case of measures, there's no one royal road to validity.

4604 You need to make a validity argument (Kane 1992).<sup>17</sup>

4605 For testing the effect of money on happiness, our manipulation was to

4606 give participants \$1,000. This manipulation is clearly face valid. But

4607 how often do people just receive a windfall of cash, versus getting a

4608 raise at work or inheriting money from a relative? Is the effect caused

4609 by *having* the money, or *receiving* the money with no strings attached?

4610 We might have to do more experiments to figure out what aspect of

4611 the money manipulation was most important. Even in straightforward

4612 cases like this one, we need to be careful about the breadth of the claims

4613 we make.

4614 Sometimes validity arguments are made based on the success of the ma-

4615 nipulation in producing some change in the measurement. In the im-

4616 plicit theory of mind case study we began with, the stimulus contained

4617 an animated Smurf character, and the argument was that participants

4618 took the Smurf's beliefs into account in making their judgments. This

4619 stimulus choice seems surprising—not only would participants have to

4620 track the implicit beliefs of other *people*, but they would also have to

4621 be tracking the beliefs of depictions of nonhuman, animated characters.

4622 On the other hand, based on the success of the manipulation, the authors

<sup>17</sup> One caveat is that the validity of a manipulation incorporates the validity of the manipulation *and* the measure. You can't really have a good estimate of a causal effect if the measurement is invalid.

4623 made an *a fortiori* argument: if people track even an animated Smurf's  
4624 beliefs, then they *must* be tracking the beliefs of real humans.

4625 Let's look at one last example to think more about manipulation validity.

4626 Walton and Cohen (2011) conducted a short intervention in which col-  
4627 lege students ( $N = 92$ ) read about social belonging and the challenges of  
4628 the transition to college and then reframed their own experiences using  
4629 these ideas. This intervention led to long-lasting changes in grades and  
4630 well-being. While the intervention undoubtedly had a basis in theory,  
4631 part of our understanding of the validity of the intervention comes from  
4632 its efficacy: sense of belonging *must* be a powerful factor if intervening  
4633 on it causes such big changes in the outcome measures.<sup>18</sup> The only dan-  
4634 ger is when the argument becomes circular—a theory is correct because  
4635 the intervention yielded a success, and the intervention is presumed to  
4636 be valid because of the theory. The way out of this circle is through  
4637 replication and generalization of the intervention. If the intervention  
4638 repeatably produces the outcome, as has been shown in replications of  
4639 the sense of belonging intervention (Walton, Brady, and Crum 2020),  
4640 then the manipulation becomes an intriguing target for future theories.  
4641 The next step in such a research program is to understand the limitations  
4642 of such interventions (sometimes called **boundary conditions**).

<sup>18</sup> On the other hand, if the manip-  
ulation *doesn't* produce a change in your  
measure, maybe the manipulation is in-  
valid, but the construct still exists. Sense  
of belonging could still be important  
even if my particular intervention failed  
to alter it!

### 4643 9.3 Summary: Experimental design

4644 In this chapter, we started by examining some common experimental  
4645 designs that allow us to measure effects associated with one or more  
4646 manipulations. Our advice, in brief, was: “Keep it simple!” The failure  
4647 mode of many experiments is that they contain too many manipulations,  
4648 and these manipulations are measured with too little precision.

4649 Start with just a single manipulation and measure it carefully. Ideally  
4650 this measurement should be done via a within-participants design un-  
4651 less the manipulation is completely incompatible with this design. And  
4652 if this design can incorporate a dose-response manipulation, it is more  
4653 likely to provide a basis for quantitative theorizing.

4654 How do you ensure that your manipulation is valid? A careful experi-  
4655 menter needs to consider possible confounds and ensure that these are  
4656 controlled or randomized. They must also consider other artifacts in-  
4657 cluding placebo, demand, and expectancy effects. Finally, they must  
4658 begin thinking about the relation of their manipulation to the broader  
4659 theoretical construct whose causal role they hope to test.



## DISCUSSION QUESTIONS

1. Choose a classic study in your area of psychology. Analyze the design choices: How many factors were manipulated? How many measures were taken? Did it use a within-participants or between-participants design? Were measures repeated? Can you justify these choices with respect to trade-offs (e.g., carryover effects, fatigue, or others)?
2. Consider the same study. Design an alternative version that varies one of these design parameters (e.g., drops a manipulation or measure or changes within- to between-participants). What are the pros and cons of this change? Do you think your design improves on the original?

4660



## READINGS

- Much of this material is covered in more depth in the classic text on research methods: Rosenthal, Robert, and Ralph L. Rosnow (2008). *Essentials of Behavioral Research: Methods and Data Analysis*. Third Edition. McGraw-Hill. <http://dx.doi.org/10.34944/dspace/66>.

4661

## 4662 References

- Baddeley, Alan D., Neil Thomson, and Mary Buchanan. 1975. “Word Length and the Structure of Short-Term Memory.” *Journal of Verbal Learning and Verbal Behavior* 14 (6): 575–589.
- Benedetti, Fabrizio. 2020. *Placebo Effects*. Oxford University Press.
- Boyce, Veronica, Maya B. Mathur, and Michael C. Frank. 2023. “Eleven

4663

- Years of Student Replication Projects Provide Evidence on the Correlates of Replicability in Psychology.” *Royal Society Open Science* 10 (11): 231240.
- Brennan, Wendy M., Elinor W. Ames, and Ronald W. Moore. 1966. “Age Differences in Infants’ Attention to Patterns of Different Complexities.” *Science* 151 (3708): 354–356.
- Clark, Herbert H. 1973. “The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research.” *Journal of Verbal Learning and Verbal Behavior* 12 (4): 335–359.
- Cunningham, Scott. 2021. *Causal Inference*. Yale University Press.
- El Kaddouri, Rachida, Lara Bardi, Diana De Bremaecker, Marcel Brass, and Roeljan Wiersema. 2020. “Measuring Spontaneous Mentalizing with a Ball Detection Task: Putting the Attention-Check Hypothesis by Phillips and Colleagues (2015) to the Test.” *Psychological Research* 84 (6): 1749–1757.
- Foroughi, Cyrus K., Samuel S. Monfort, Martin Paczynski, Patrick E. McKnight, and P. M. Greenwood. 2016. “Placebo Effects in Cognitive Training.” *Proceedings of the National Academy of Sciences* 113 (27): 7470–7474.
- Gelman, Andrew. 2017. “Poisoning the Well with a Within-Person Design? What’s the Risk?” In *Statistical Modeling, Causal Inference, and Social Science*. <https://statmodeling.stat.columbia.edu/2017/11/25/poisoning-well-within-person-design-whats-risk/>.
- Greenwald, Anthony G. 1976. “Within-Subjects Designs: To Use or Not to Use?” *Psychological Bulletin* 83 (2): 314–320.
- Jaeggi, Susanne M., Martin Buschkuhl, John Jonides, and Walter J. Perrig. 2008. “Improving Fluid Intelligence with Training on Working Memory.”

- Proceedings of the National Academy of Sciences* 105 (19): 6829–6833.
- Kane, Michael T. 1992. “An Argument-Based Approach to Validity.” *Psychological Bulletin* 112 (3): 527–535.
- Kidd, Celeste, Steven T. Piantadosi, and Richard N. Aslin. 2012. “The Goldilocks Effect: Human Infants Allocate Attention to Visual Sequences That Are Neither Too Simple Nor Too Complex.” *PLoS One* 7 (5): e36399.
- Kovács, Ágnes Melinda, Ernő Téglás, and Ansgar Denis Endress. 2010. “The Social Sense: Susceptibility to Others’ Beliefs in Human Infants and Adults.” *Science* 330 (6012): 1830–1834.
- Lakens, Daniel. 2016. “Why Within-Subject Designs Require Fewer Participants than Between-Subject Designs.” 2016. <https://daniellakens.blogspot.com/2016/11/why-within-subject-designs-require-less.html>.
- Lovatt, Peter, Steve E. Avons, and Jackie Masterson. 2000. “The Word-Length Effect and Disyllabic Words.” *The Quarterly Journal of Experimental Psychology: Section A* 53 (1): 1–22.
- McClelland, Gary H., and Charles M. Judd. 1993. “Statistical Difficulties of Detecting Interactions and Moderator Effects.” *Psychological Bulletin* 114 (2): 376–390.
- Melby-Lervåg, Monica, and Charles Hulme. 2013. “Is Working Memory Training Effective? A Meta-Analytic Review.” *Developmental Psychology* 49 (2): 270–291.
- Myung, Jay I., and Mark A. Pitt. 2009. “Optimal Experimental Design for Model Discrimination.” *Psychological Review* 116 (3): 499–518.
- Orne, Martin T. 1962. “On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their

- Implications.” *American Psychologist* 17 (11): 776–783.
- Phillips, Jonathan, Desmond C. Ong, Andrew D. R. Surtees, Yijing Xin, Samantha Williams, Rebecca Saxe, and Michael C. Frank. 2015. “A Second Look at Automatic Theory of Mind: Reconsidering Kovács, Téglás, and Endress (2010).” *Psychological Science* 26 (9): 1353–1367.
- Redick, Thomas S., Zach Shipstead, Tyler L. Harrison, Kenny L. Hicks, David E. Fried, David Z. Hambrick, Michael J. Kane, and Randall W. Engle. 2013. “No Evidence of Intelligence Improvement after Working Memory Training: A Randomized, Placebo-Controlled Study.” *Journal of Experimental Psychology: General* 142 (2): 359–379.
- Rosenthal, Robert. 1994. “Interpersonal Expectancy Effects: A 30-Year Perspective.” *Current Directions in Psychological Science* 3 (6): 176–179.
- Rosenthal, Robert, and Ralph L. Rosnow. 2008. *Essentials of Behavioral Research: Methods and Data Analysis*. 3rd ed. McGraw-Hill. <https://doi.org/10.34944/dspace/66>.
- Rosnow, Ralph, and Robert Rosenthal. 1997. *People Studying People: Artifacts and Ethics in Behavioral Research*. W. H. Freeman.
- Simons, Daniel J., Walter R. Boot, Neil Charness, Susan E. Gathercole, Christopher F. Chabris, David Z. Hambrick, and Elizabeth A. L. Stine-Morrow. 2016. “Do ‘Brain-Training’ Programs Work?” *Psychological Science in the Public Interest* 17 (3): 103–186.
- Walton, Gregory M., Shannon T. Brady, and A. J. Crum. 2020. “The Social-Belonging Intervention.” *Handbook of Wise Interventions: How Social Psychology Can Help People Change*, 36–62.
- Walton, Gregory M., and Geoffrey L. Cohen. 2011. “A Brief Social-

Belonging Intervention Improves Academic and Health Outcomes of Minority Students.” *Science* 331 (6023): 1447–1451.

Westfall, Jacob, Charles M. Judd, and David A. Kenny. 2015. “Replicating Studies in Which Samples of Participants Respond to Samples of Stimuli.” *Perspectives on Psychological Science* 10 (3): 390–399.

Young, Liane, Fiery Cushman, Marc Hauser, and Rebecca Saxe. 2007. “The Neural Basis of the Interaction between Theory of Mind and Moral Judgment.” *Proceedings of the National Academy of Sciences* 104 (20): 8235–8240.

## 10 SAMPLING

4668



### LEARNING GOALS

- Discuss sampling theory and stratified sampling
- Reason about the limitations of different samples, especially convenience samples
- Consider sampling biases and how they affect your inferences
- Learn how to choose and justify an appropriate sample size for your experiment

4669

4670 As we keep reminding you, experiments are designed to yield measure-  
4671 ments of a causal effect. But a causal effect of what, and for whom?  
4672 These are questions that are often given surprisingly little air time in  
4673 our papers. Titles in our top journals read “Daxy thinking promotes  
4674 fribbles,” “Doing fonzy improves smoodling,” or “Blicket practice pro-  
4675 duces more foozles than smonkers.”<sup>1</sup> Each of these uses **generic lan-**  
4676 **guage** to state a claim that is implied to be generally true (DeJesus et al.  
4677 2019),<sup>2</sup> but for each of these, we could reasonably ask, “For whom?” Is

<sup>1</sup> Titles changed to protect the original authors. These researchers might very well have said more specific things in the text of their paper.

<sup>2</sup> Generic language is a fascinating linguistic phenomenon. When we say things like “Mosquitoes transmit malaria,” we don’t mean that *all* mosquitoes do it, only something like “It’s a valid and diagnostic generalization about mosquitoes in contrast to other relevant insects or other creatures that they are spreaders of malaria” (see Tessler and Goodman 2019).

4678 it everyone? Or a particular set of people? These are questions about  
4679 our key theme, GENERALIZABILITY.

4680 Let's focus on smoodling. We wouldn't let the authors get away with  
4681 a fully universal version of their claim: "Doing [*any*] fonzy improves  
4682 smoodling [*for everyone*]." The nongeneric version states a generalization  
4683 that goes way beyond the evidence we actually have. But it seems that  
4684 we are often okay with authors *implying* (with generic language) that  
4685 their findings generalize broadly. Imagine for a second what the com-  
4686 pletely specific version of one of these titles might look like: "Reading  
4687 one particular selection of fonzy for fifteen minutes in the lab improved  
4688 36 college students' smoodling scores on a questionnaire." This paper  
4689 sounds pretty narrow in its applicability!

4690 We've already run into generalizability in our treatment of statistical  
4691 estimation and inference. When we estimated a particular quantity (say,  
4692 the effect of fonzy), we did so in our own sample. But we then used  
4693 inferential tools to reason about how the estimate in this **sample** related  
4694 to the parameter in the **population** as a whole. How do we link up  
4695 these *statistical* tools for generalization to the *scientific* questions we have  
4696 about the generalizability of our findings? That's the question of this  
4697 chapter.

4698 A key set of decisions in experiment planning is what population to

4699 sample from and how to sample. We'll start by talking about the basics  
4700 of **sampling theory**: different ways of sampling and the generalizations  
4701 they do and don't license. The second section of the chapter will then  
4702 deal with **sampling biases** that can compromise our effect estimates. A  
4703 final set of key decisions is about **sample size planning**. In the third part  
4704 of the chapter we'll address this issue, starting with classic **power analysis**  
4705 but then introducing several other ways that an experimenter can plan  
4706 and justify their sample size.

## 4707 *10.1 Sampling theory*

4708 The basic idea of sampling is simple: you want to estimate some  
4709 measurement for a large or infinite population by measuring a sample  
4710 from that population.<sup>3</sup> Sampling strategies are split into two categories.  
4711 **Probability sampling** strategies are those in which each member of the  
4712 population has some known, prespecified probability of being selected  
4713 to be in the sample—think “generalizing to Japanese people by picking  
4714 randomly from a list of everyone in Japan.” **Non-probability sampling**  
4715 covers strategies in which probabilities are unknown or shifting, or in  
4716 which some members of the population could never be included in  
4717 the sample—think “generalizing to Germans by sending a survey to  
4718 a German email list and asking people to forward the email to their

<sup>3</sup> There are some tools for dealing with estimation in smaller populations where your sample is a substantial fraction of the population (e.g., a survey of your department where you get responses from half of the students). We won't discuss those here; our focus is on generalizing to large populations of humans.

<sup>4719</sup> family.”



### CASE STUDY

#### *Is everyone bad at describing smells?*

Since Darwin, scientists have assumed that smell is a vestigial sense in humans—one that we don’t even bother to encode in language. In English we don’t even have consistent words for odors. We can say something is “stinky,” “fragrant,” or maybe “musty,” but beyond these, most of our words for smells are about the *source* of the smell, not the qualities of it. Bananas, roses, and skunks all have distinctive smells, but we don’t have any vocabulary for naming what is common or uncommon about them. And when we make up ad hoc vocabulary, it’s typically quite inconsistent (Majid and Burenhult 2014). The same situation applies across many languages.

So, would it be a good generalization about human beings—all people—that olfaction as a sense is de-emphasized relative to, say, vision? This inference has a classic sample-to-population structure. Within several samples of participants using widely spoken languages, we observe limited and inconsistent vocabulary for smells, as well as poor discrimination. We use these samples to license an inference to the population—in this case, the entire human population.

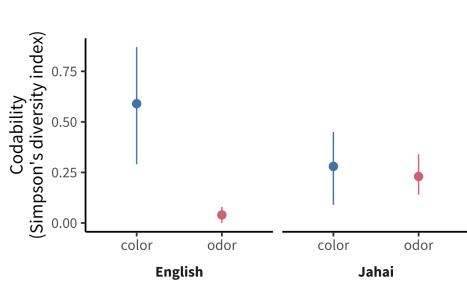


Figure 10.1

Data from Majid and Burenhult (2014) on the consistency of color and odor naming in English and Jahai speakers. Higher values indicate more consistent descriptions. Error bars show standard deviation.

But these inferences about the universal lack of olfactory vocabulary are likely based on choosing nonrepresentative samples! Multiple hunter-gatherer groups appear to have large vocabularies for consistent smell description. For example, the Jahai, a hunter-gatherer group on the Malay Peninsula, have a vocabulary that includes at least twelve words for distinct odors, for example /cŋɛs/, which names odors with a “stinging smell” like gasoline, smoke, or bat droppings. When Jahai speakers are asked to name odors, they produce shorter and much more consistent descriptions than English speakers—in fact, their smell descriptions were as consistent as their color descriptions (figure 10.1). Further studies implicate the hunter-gatherer lifestyle as a factor: while several hunter-gatherer groups show good odor naming, nearby horticulturalist groups don’t (Majid and Kruspe 2018).

Generalizations about humans are tricky. If you want to estimate the average odor naming ability, you could take a random sample of humans and evaluate their odor naming. Most of the individuals in the sample would likely speak English, Mandarin, Hindi, or Spanish. Almost cer-

tainly, none of them would speak Jahai, which is spoken by only a little more than a thousand people and is listed as “endangered” by Ethnologue (<https://www.ethnologue.com/language/jhi>). Your estimate of low odor naming stability might be a good guess for the *majority* of the world’s population, but would tell you little about the Jahai.

On the other hand, it’s more complicated to jump from a statistical generalization about average ability to a richer claim like, “Humans have low olfactory naming ability.” Such claims about universal aspects of the human experience require much more care and much stronger evidence (Pi-antadosi and Gibson 2014). From a sampling perspective, human behavior and cognition show immense and complex heterogeneity—variability of individuals and variability across clusters. Put simply, if we want to know what people in general are like, we have to think carefully about which people we include in our studies.

4722

#### 4723 10.1.1 Classical probability sampling

4724 In classical sampling theory, there is some **sampling frame** containing ev-  
ery member of the population—think of a giant list with every adult hu-  
4725 man’s name in it. Then we use some kind of **sampling strategy**, maybe  
4726 at the simplest just a completely random choice, to select  $N$  humans  
4727 from that sample frame, and then we include them in our experiment.

4728 This scenario is the one that informs all of our statistical results about

4730 how sample means converge to the population mean (as in chapter 6).

4731 Unfortunately, we very rarely do sampling of this sort in psychological

4732 research. Gathering true probability samples from the large populations

4733 that we'd like to generalize to is far too difficult and expensive. Con-

4734 sider the problems involved in doing some experiment with a sample

4735 of *all adult humans*, or even *adult English-speaking humans who are located*

4736 *in the United States*. As soon as you start to think about what it would

4737 take to collect a probability sample of this kind of population, the com-

4738 plexities get overwhelming. How will you find their names—what if

4739 they aren't in the phone book? How will you contact them—what if

4740 they don't have email? How will they do your experiment—what if

4741 they don't have an up-to-date web browser? What if they don't want

4742 to participate at all?

4743 Instead, the vast majority of psychology research has been conducted

4744 with **convenience samples**: non-probability samples that feature indi-

4745 viduals who can be recruited easily, such as college undergraduates or

4746 workers on crowdsourcing platforms like Amazon Mechanical Turk or

4747 Prolific Academic (see chapter 12). We'll turn to these below.

4748 For survey research, on the other hand—think of election polling—there

4749 are many sophisticated techniques for dealing with sampling; although

4750 this field is still imperfect, it has advanced considerably in trying to pre-  
4751 dict complex and dynamic behaviors. One of the basic ideas is the con-  
4752 struction of **representative samples**: samples that resemble the popula-  
4753 tion in their representation of one or several sociodemographic charac-  
4754 teristics like gender, income, race and ethnicity, age, or political ori-  
4755 entation.

4756 Representative samples can be constructed by probability sampling, but  
4757 they can also be constructed through non-probability methods like re-  
4758 cruiting quotas of individuals from different groups via various different  
4759 convenience methods. These methods are critical for much social sci-  
4760 ence research, but they have been used less frequently in experimental  
4761 psychology research and aren't necessarily a critical part of the begin-  
4762 ning experimentalist's toolkit.<sup>4</sup>

<sup>4</sup> Readers can come up with counter-examples of recent studies that focus on representative sampling, but our guess is that they will prove the rule more generally. For example, a recent study tested the generality of growth mindset interventions for US high school students using a national sample (Yeager et al. 2019). This large-scale study sampled more than 100 high schools from a sampling frame of all registered high schools in the US, then randomly assigned students within schools that agreed to participate. They then checked that the schools that agreed to participate were representative of the broader population of schools. This study is great stuff, but we hope you agree that if you find yourself in this kind of situation—planning a multi-investigator five-year consortium study on a national sample—you might want to consult with a statistician and not use an introductory book like this one.

## DEPTH

### *Representative samples and stratified sampling*

Stratified sampling is a cool method that can help you get more precise estimates of your experimental effect, if you think it varies across some grouping in your sample. Imagine you're interested in a particular measure in a population—say, attitudes toward tea drinking across US adults—but you think that this measure will vary with one or more characteristics such as whether the adults are frequent, infrequent, or non-coffee

drinkers. Even worse, your measure might be more variable within one group: perhaps most frequent and infrequent coffee drinkers feel okay about tea, but as a group, non-coffee drinkers tend to hate it (most don't drink any caffeinated beverages).

A simple random sample from this heterogeneous population *will* yield statistical estimates that converge asymptotically to the correct population average for tea-drinking attitudes. But it will do so more slowly than ideal because any given sample may over- or under-sample nondrinkers just by chance. In a small sample, if you happen to get too many non-coffee drinkers, your estimate of attitudes will be biased downward; if you happen to get too few, you will be biased upward. All of this will come out in the wash eventually, but any individual sample (especially a small one) will be noisier than ideal.

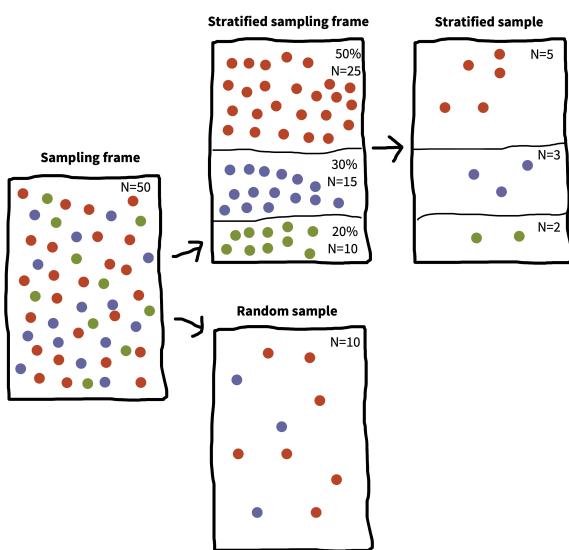


Figure 10.2

An illustration of stratified sampling. The left panel shows the sampling frame. The upper frames show the sampling frame stratified by a participant characteristic and a stratified sample. The lower frame shows a simple random sample, which happens to omit one group completely by chance.

But, if you know the proportion of frequent, infrequent, or non-coffee drinkers in the population, you can perform stratified sampling within those subpopulations to ensure that your sample is representative along this dimension ([Neyman 1992](#)). This situation is pictured in figure 10.2, which shows how a particular sampling frame can be broken up into groups for stratified sampling. The result is a sample that matches the population proportions on a particular characteristic. In contrast, a simple random sample can over- or under-sample the subgroups by chance.

Stratified sampling can lead to substantial gains in the precision of your estimate. These gains are most prominent when either the groups differ a lot in their mean or when they differ a lot in their variance. There are several important refinements of stratified sampling in case you think these methods are important for your problem. In particular, **optimal sampling** can help you figure out how to over-sample groups with higher variance. On the other hand, if the characteristic on which you stratify participants doesn't relate to your outcome at all, then estimates from stratified sampling converge just as fast as random sampling (though it's a bit more of a pain to implement).

Figure 10.3 shows a simulation of the scenario in figure 10.2, in which each coffee preference group has a different tea attitude mean and the smallest group has the biggest variance. Although the numbers here are invented, it's clear that estimation error is much smaller in the stratified group and estimation error declines much more quickly as samples get larger.

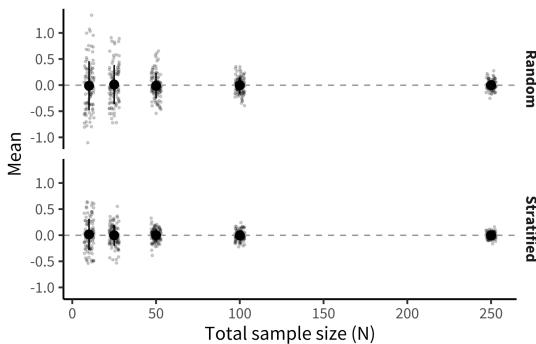


Figure 10.3

A simulation showing the potential benefits of stratification. Each dot is an estimated mean for a sample of a particular size, sampled randomly or with stratification. Red points show the mean and standard deviation of sample estimates.

Stratification is everywhere, and it's useful even in convenience samples.

For example, researchers who are interested in development typically stratify their samples across ages (e.g., recruiting equal numbers of two- and three-year-olds for a study of preschoolers). You can estimate developmental change in a pure random sample, but you are guaranteed good coverage of the range of interest when you stratify.

If you have an outcome that you think varies with a particular characteristic, it's not a bad idea to consider stratification. But don't go overboard—you can drive yourself to distraction finding the last left-handed nonbinary coffee drinker to complete your sample. Focus on stratifying when you know the measure varies with the characteristic of interest.

4767 10.2 *Convenience samples, generalizability, and the*

4768 *WEIRD problem*

4769 Now let's go back to the question of generalizability. How generalizable

4770 are the experimental effect estimates that we obtain in experiments that

4771 are conducted only with convenience samples? We'll start by laying

4772 out the worst version of the problem of generalizability in experimental

4773 psychology. We'll then try to pull back from the brink and discuss some

4774 reasons why we might not want to be in despair despite some of the

4775 generalizability issues that plague the psychology literature.

4776 10.2.1 *The worst version of the problem*

4777 Psychology is the study of the human mind. But from a sampling theory

4778 standpoint, not a single estimate in the published literature is based on a

4779 simple random sample from the human population. And the situation

4780 is worse than that. Here are three of the most severe issues that have

4781 been raised regarding the generalizability of psychology research.

4782 1. **Convenience samples.** Almost all research in experimental psy-

4783 chology is performed with convenience samples. This issue has

4784 led to the remark that “the existing science of human behavior

4785 is largely the science of the behavior of sophomores” ([McNemar](#)

4786 1946, p. 333; quoted in Rosenthal and Rosnow 1984, p. 261).

4787 The samples we have easy access to just don't represent the pop-  
4788 ulations we want to describe! At some point there was a social  
4789 media account devoted to finding biology papers that made big  
4790 claims about curing diseases and appending the qualifier "in mice"  
4791 to them. We might consider whether we need to do the same to  
4792 psychology papers. Would "Doing *fonzy improves smoodling* *in*  
4793 *sophomore college undergraduates in the Western US*" make it into a  
4794 top journal?

4795 2. **The WEIRD problem.** Not only are the convenience samples  
4796 that we study not representative of the local or national con-  
4797 texts in which they are recruited but those local and national  
4798 contexts are also unrepresentative of the broad range of human  
4799 experiences. Henrich, Heine, and Norenzayan (2010) coined  
4800 the term WEIRD (Western, educated, industrialized, rich, and  
4801 democratic) to sum up some of the ways that typical participants  
4802 in psychology experiments differ from other humans. The vast  
4803 over-representation of WEIRD participants in the literature  
4804 has led some researchers to suggest that published results simply  
4805 reflect "WEIRD psychology"—a small and idiosyncratic part of  
4806 a much broader universe of human psychology.<sup>5</sup>

5 The term WEIRD has been very useful in drawing attention to the lack of representation of the breadth of human experiences in experimental psychology. But one negative consequence of this idea has been the response that what we need to do as a field is to sample more "non-WEIRD" people. It is not helpful to suggest that every culture outside the WEIRD moniker is the same (Syed and Kathawalla 2020). A better starting point is to consider the way that cultural variation might guide our choices about sampling.

4807     3. The item sampling issue. As we discussed in chapter 7 and 9, we're  
4808        typically not just trying to generalize to new people; we're also  
4809        trying to generalize to new stimuli (Westfall, Judd, and Kenny  
4810        2015). The problem is that our experiments often use a very small  
4811        set of items, constructed by experimenters in an ad hoc way rather  
4812        than sampled as representatives of a broader population of stim-  
4813        uli that we hope to generalize to with our effect size estimate.  
4814        What's more, our statistical analyses sometimes fail to take stimu-  
4815        lus variation into account. Unless we know about the relationship  
4816        of our items to the broader population of stimuli, our estimates  
4817        may be based on unrepresentative samples in yet another way.

4818     In sum, experiments in the psychology literature primarily measure ef-  
4819        fects from WEIRD convenience samples of people and unsystematic  
4820        samples of experimental stimuli. Should we throw up our hands and  
4821        resign ourselves to an ungeneralizable “science” of sample-specific anec-  
4822        dotes (Yarkoni 2020)?

4823     10.2.2 *Reasons for hope and ways forward*

4824     We think the situation isn't as bleak as the arguments above might have  
4825        suggested. Underlying each of the arguments above is the notion of  
4826        heterogeneity, the idea that particular effects vary in the population.

4827 Let's think through a very simple version of this argument. Say we have  
 4828 an experiment that measures the smoodling effect, and it turns out that  
 4829 smoodling is completely universal and invariant throughout the human  
 4830 population. Now, if we want to get a precise estimate of smoodling,  
 4831 we can take *any* sample we want because everyone will show the same  
 4832 pattern. Because smoodling is homogeneous, a nonrepresentative sam-  
 4833 ple will not cause problems. There are some phenomena like this! For  
 4834 example, the Stroop task produces a consistent and similar interference  
 4835 effect for almost everyone (Hedge, Powell, and Sumner 2018).

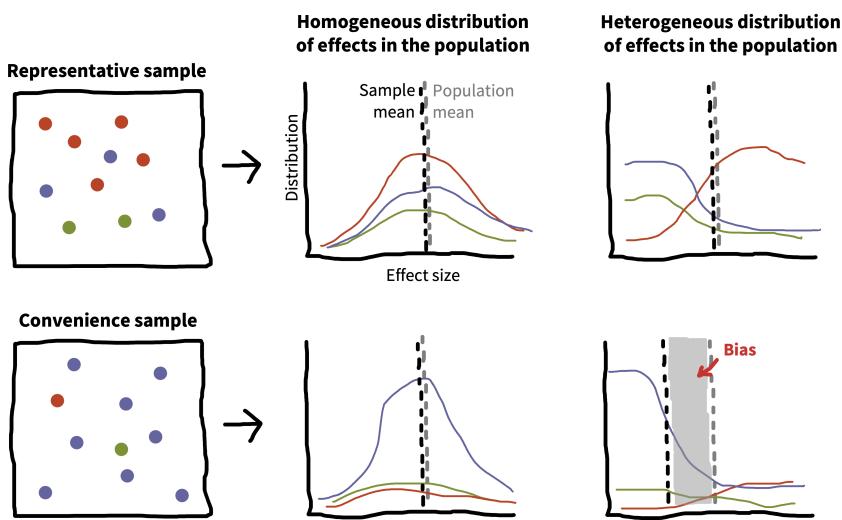


Figure 10.4

An illustration of the interaction of heterogeneity and convenience samples. Colors indicate arbitrary population subgroups. The left-hand panels show sample composition. Individual plots show the distribution of effect sizes in each subgroup.

4836 Figure 10.4 illustrates this argument more broadly. If you have a rep-  
 4837 resentative sample (top), then your sample mean and your population  
 4838 mean will converge to the same value, regardless of whether the effect  
 4839 is homogeneous (right) or heterogeneous (right). That's the beauty of  
 4840 sampling theory. If you have a convenience sample, one part of the

4841 population is overrepresented in the sample. The convenience sam-  
4842 ple doesn't cause problems if the size of your effect is homogeneous  
4843 in the population—as with the case of smoodling or Stroop. The trou-  
4844 ble comes when you have an effect that is heterogeneous. Because one  
4845 group is overrepresented, you get systematic bias in the sample mean  
4846 relative to the population mean.

4847 So the problems listed above—convenience samples, WEIRD samples,  
4848 and narrow stimulus samples—only cause issues if effects are heteroge-  
4849 neous. Are they? The short answer is, *we don't know*. Convenience  
4850 samples are fine in the presence of homogeneous effects, but we only  
4851 use convenience samples, so we may not know which effects are homo-  
4852 geneous! Our metaphorical heads are in the sand.

4853 We can't do better than this circularity without a theory of what should  
4854 be variable and what should be consistent between individuals.<sup>6</sup> As  
4855 naive observers of human behavior, differences between people often  
4856 loom large. We are keen observers of social characteristics like age, gen-  
4857 der, race, class, and education. For this reason, our intuitive theories  
4858 of psychology often foreground these characteristics as the primary lo-  
4859 cus for variation between people. Certainly these characteristics are  
4860 important, but they fail to explain many of the *invariances* of human  
4861 psychology as well. An alternative line of theorizing starts with the idea

<sup>6</sup> Many people have theorized about the ways that culture and language in general might moderate psychological processes (e.g., Markus and Kitayama 1991). What we're talking about is related but slightly different—a theory not of what's different but of when there should be any difference and when there shouldn't be. As an example, Tsai's (2007) "ideal affect" theory predicts that there should be more similarities in the distribution of actual affect across cultures, but that cultural differences should emerge in *ideal affect* (what people want to feel like) across cultures. This is a theory of when you should see homogeneity and when you should see heterogeneity.

4862 that “lower-level” parts of psychology—like perception—should be less  
4863 variable than “higher-level” faculties like social cognition. This kind  
4864 of theory sounds like a useful place to start, but there are also counter-  
4865 examples in the literature, including cases of cultural variation in per-  
4866ception (Henrich, Heine, and Norenzayan 2010).

4867 Multi-lab, multi-nation studies can help to address questions about het-  
4868 erogeneity, breaking the circularity we described above. For example,  
4869 ManyLabs 2 systematically investigated the replicability of a set of phe-  
4870 nomena across cultures (Klein et al. 2018), finding limited variation in  
4871 effects between WEIRD sites and other sites. And in a study compar-  
4872 ing a set of convenience and probability samples, Coppock, Leeper, and  
4873 Mullinix (2018) found limited demographic heterogeneity in another  
4874 sample of experimental effects from across the social sciences. So there  
4875 are at least some cases where we don’t have to worry as much about het-  
4876 erogeneity. More generally, such large-scale studies offer the possibility  
4877 of measuring and characterizing demographic and cultural variation—  
4878 as well as how variation itself varies between phenomena.

### 4879 *10.3 Biases in the sampling process*

4880 In fields like econometrics or epidemiology that use observational meth-  
4881 ods to estimate causal effects, reasoning about **sampling biases** is a critical

part of estimating generalizable effects. If your sample does not represent the population of interest, then your effect estimates will be biased.<sup>7</sup>

In the kind of experimental work we are discussing, many of these issues are addressed by random assignment, including the first issue we treat: **collider bias**. Not so for the second one, **attrition bias**, which is an issue even in randomized experiments.

### 10.3.1 Collider bias

Imagine you want to measure the association between money and happiness through a (nonexperimental) survey. As we discussed in chapter 1, there are plenty of causal processes that could lead to this association. Figure 10.5 shows several of these scenarios. Money could truly cause happiness (1); happiness could cause you to make more money (2); or some third factor—say having lots of friends—could cause people to be happier *and* richer (3).

But we can also create spurious associations if we are careless in our sampling. One prominent problem that we can induce is called **collider bias**. Suppose we recruited our sample from the clients of a social services agency. Unfortunately, both of our variables might affect presence in a social service agency (figure 10.5, 4): people might be interacting

<sup>7</sup> There is a deep literature on correcting these biases using causal inference frameworks. These techniques are well outside of the scope of this book, but if you’re interested, you might look at some of the textbooks we recommended earlier, such as Cunningham (2021).

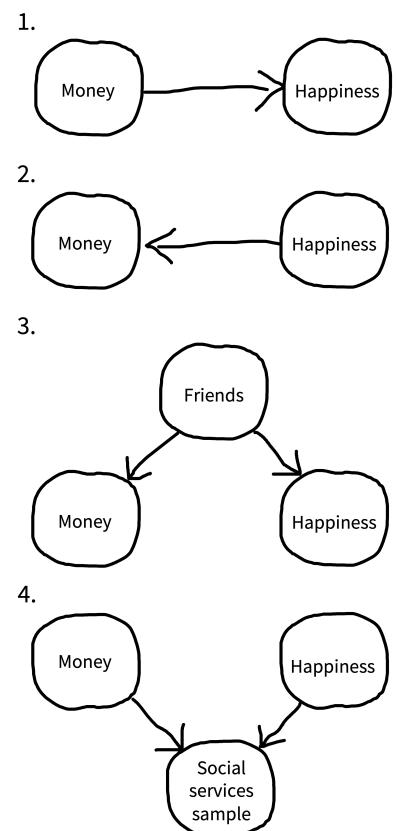


Figure 10.5  
Four reasons why money and happiness can be correlated in a particular sample: (1) causal relationship, (2) reverse causality, (3) confounding with friendship, and (4) collider bias. For this last scenario, we have to assume that our measurement is *conditioned* on being in this sample, meaning we only look at the association of money and happiness within the social services sample.

4901 with the agency for financial or benefits assistance, or else for psycho-  
4902 logical services (perhaps due to depression).

4903 Being in a social services sample is called a **collider** variable because the  
4904 two causal arrows *collide* into it (they both point to it). If we look just  
4905 within the social services sample, we might see a *negative* association  
4906 between wealth and happiness—on average the people coming for fi-  
4907 nancial assistance would have less wealth and more happiness than the  
4908 people coming for psychological services. The take-home here is that  
4909 in observational research, you need to think carefully about the causal  
4910 structure of your sampling process (Rohrer 2018).

4911 If you are doing experimental research, you are mostly protected from  
4912 this kind of bias: Random assignment still “works” even in subselected  
4913 samples. If you run a money intervention within a social-services popu-  
4914 lation using random assignment, you can still make an unbiased estimate  
4915 of the effect of money on happiness. But that estimate will only be valid  
4916 *for members of that subselected population.*

#### 4917 10.3.2 Attrition bias

4918 **Attrition** is when people drop out of your study. You should do every-  
4919 thing you can to improve participants’ experiences (see chapter 12), but  
4920 sometimes—especially when a manipulation is onerous for participants

4921 or your experiment is longitudinal and requires tracking participants for  
 4922 some time—you will still have participants withdraw from the study.

4923 Attrition on its own can be a threat to the generalizability of an experi-  
 4924 mental estimate. Imagine you do an experiment comparing a new very  
 4925 intense after-school math curriculum to a control curriculum in a sam-  
 4926 ple of elementary school children over the course of a year. By the end  
 4927 of the year, suppose many of your participants have dropped out. The  
 4928 families who have stayed in the study are likely those who care most  
 4929 about math. Even if you see an effect of the curriculum intervention,  
 4930 this effect may generalize only to children in families who love math.

4931 But there is a further problem with attrition, known as **selective attrition**.  
 4932 If attrition is related to the outcome specifically within the treatment  
 4933 group (or for that matter, specifically within the control group), you can  
 4934 end up with a biased estimate, even in the presence of random assign-  
 4935 ment ([Nunan, Aronson, and Bankhead 2018](#)). Imagine that students in  
 4936 the control condition of your math intervention experiment stayed in  
 4937 the sample, but the math intervention itself was so tough that most fam-  
 4938 ilies dropped out except those who were very interested in math. Now,  
 4939 when you compare math scores at the end of the experiment, your esti-  
 4940 mate will be biased (figure 10.6): scores in the math condition could be  
 4941 higher simply because of differences in who stuck around to the end.<sup>8</sup>

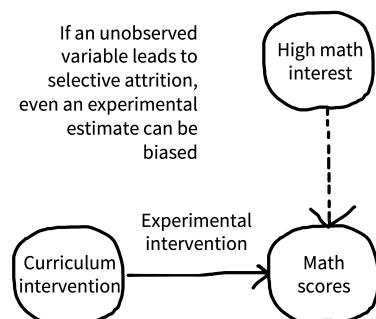


Figure 10.6  
 Selective attrition can lead to a bias even in the presence of random assignment. The dashed line indicates a causal relationship that is unobserved by the researcher.

<sup>8</sup> If you get deeper into drawing DAGs like we are doing here, you will want to picture attrition as its own node in the graph, but that's beyond the scope of this book.

4942 Unfortunately, it turns out that attrition bias can be pretty common  
4943 even in short studies, especially when they are conducted online when  
4944 a participant can drop out simply by closing a browser window. This  
4945 bias can be serious enough to lead to false conclusions. For example,  
4946 Zhou and Fishbach (2016) ran an experiment in which they asked on-  
4947 line participants to write about either four happy events (low difficulty)  
4948 or 12 happy events (high difficulty) from the last year and then asked  
4949 the participants to rate the difficulty of the task. Surprisingly, the high-  
4950 difficulty task was rated as easier than the low-difficulty task! Selective  
4951 attrition was the culprit for this counterintuitive conclusion: while only  
4952 26% of participants dropped out of the low-difficulty condition, a full  
4953 69% dropped out of the high-difficulty task. The 31% that were left  
4954 found it quite easy for them to generate 12 happy events, and so they  
4955 rated the objectively harder task as less difficult.

4956 Always try to track and report attrition information. That lets you—and  
4957 others—understand whether attrition is leading to bias in your estimates  
4958 or threats to the generalizability of your findings.<sup>9</sup>

#### 4959 10.4 Sample size planning

4960 Now that you have spent some time considering your sample and what  
4961 population it represents, how many people will your sample contain?

<sup>9</sup> If you get interested, there is a whole field of statistics that focuses on **missing data** and provides models for reasoning about and dealing with cases where data might not be **missing completely at random** (Little and Rubin 2019 is the classic reference for these tools). The causal inference frameworks referenced above also have very useful ways of thinking about this sort of bias.

4962 Continuing to collect data until you observe a  $p < 0.05$  in an inferen-  
4963 tial test is a good way to get a false positive. This practice, known as  
4964 “optional stopping,” is a good example of a practice that invalidates  $p$ -  
4965 values, much like the cases of analytic flexibility discussed in chapter 3  
4966 and chapter 6.

4967 Decisions about when to stop collecting data should not be data-  
4968 dependent. Instead you should transparently declare your data  
4969 collection **stopping rule** in your study preregistration (see chapter 11).  
4970 This step will reassure readers that there is no risk of bias from optional  
4971 stopping. The simplest stopping rule is, “I’ll collect data until I get to a  
4972 target  $N$ ”—all that’s needed in this case is a value for  $N$ .

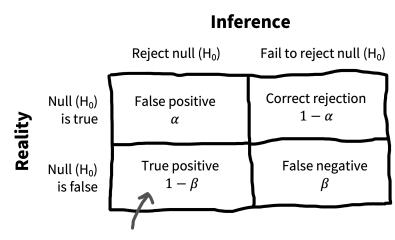
4973 But how do you decide  $N$ ? It’s going to be dependent on the effect  
4974 that you want to measure and how it varies in the population. Smaller  
4975 effects will require larger sample sizes. Classically,  $N$  was computed  
4976 using **power analysis**, which can provide a sample size for which you  
4977 have a good chance of rejecting the null hypothesis (given a particular  
4978 expected effect size). We’ll introduce this computation below.

4979 Classical power analysis is not the only way to plan your sample size.  
4980 There are a number of other useful strategies, some of which rely on  
4981 the same kinds of computations as power analysis (table 10.1). Each of

- <sup>4982</sup> these can provide a valid justification for a particular sample size, but  
<sup>4983</sup> they are useful in different situations.

Table 10.1  
 Types of data collection stopping rules.

Method	Stopping rule	Example
Power analysis	Stop at N for known probability of rejecting the null given known effect size	Randomized trial with strong expectations about effect size
Resource constraint	Stop collecting data after a certain amount of time or after a certain amount of resources are used	Time-limited field work
Smallest effect size of interest	Stop at N for known probability of rejecting the null for effects greater than some minimum	Measurement of a theoretically important effect with unknown magnitude
Precision analysis	Stop at N that provides some known degree of precision in measure	Experimental measurement to compare with predictions of cognitive models
Sequential analysis	Stop when a known inferential criterion is reached	Intervention trial designed to accept or reject null with maximal efficiency



Power to reject the null

Figure 10.7  
 A standard decision matrix for null-hypothesis significance testing.

4984 10.4.1 Power analysis

4985 Let's start by reviewing the null-hypothesis significance testing  
4986 paradigm that we introduced in chapter 6. Recall that we introduced  
4987 the Neyman-Pearson decision-theoretic view of testing in chapter 6,  
4988 shown again in figure 10.7. The idea was that we've got some null  
4989 hypothesis  $H_0$  and some alternative  $H_1$ —something like “No effect”  
4990 and “Yes, there is some effect with known size”—and we want to use  
4991 data to decide which state we're in.  $\alpha$  is our criterion for rejecting the  
4992 null, conventionally set to  $\alpha = 0.05$ .

4993 But what if  $H_0$  is actually false and the alternative  $H_1$  is true? Not  
4994 all experiments are equally well set up to reject the null in those cases.  
4995 Imagine doing an experiment with  $N = 3$ . In that case, we'd almost  
4996 always fail to reject the null, even if it were false. Our sample would  
4997 almost certainly be too small to rule out sampling variation as the source  
4998 of our observed data.

4999 Let's try to quantify our willingness to miss the effect—the false negative  
5000 rate. We'll denote this probability with  $\beta$ . If  $\beta$  is the probability of  
5001 missing an effect (failing to reject the null when it's really false), then  
5002  $1 - \beta$  is the probability that we correctly reject the null when it is false. That's  
5003 what we call the **statistical power** of the experiment.

5004 We can only compute power if we know the effect size for the alterna-  
 5005 tive hypothesis. If the alternative hypothesis is a small effect, then the  
 5006 probability of rejecting the null will typically be low (unless the sample  
 5007 size is very large). In contrast, if the alternative hypothesis is a large  
 5008 effect, then the probability of rejecting the null will be higher.

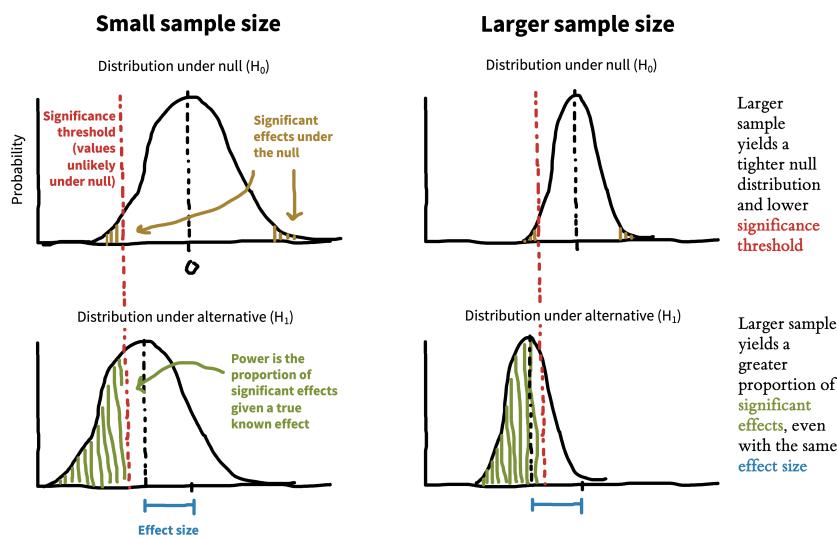


Figure 10.8  
 An illustration of how larger sample sizes lead to greater power.

5009 The same dynamic holds with sample size: the same effect size will be  
 5010 easier to detect with a larger sample size than with a small one. Fig-  
 5011 ure 10.8 shows how this relationship works. A large sample size cre-  
 5012 ates a tighter null distribution (right side) by reducing sampling error.  
 5013 A tighter null distribution means you can reject the null more of the  
 5014 time based on the variation in a true effect. If your sample size is too  
 5015 small to detect your effect much of the time, we call this being under-  
 5016 powered.<sup>10</sup>

<sup>10</sup> You can also refer to a design as over-powered, though we object slightly to this characterization, since the value of large datasets is typically not just to reject the null but also to measure an effect with high precision and to investigate how it is moderated by other characteristics of the sample.

5017 Classical power analysis involves computing the sample size  $N$  that's  
5018 necessary in order to achieve some level of power, given  $\alpha$  and a known  
5019 effect size.<sup>11</sup> The mathematics of the relationship between  $\alpha$ ,  $\beta$ ,  $N$ ,  
5020 and effect size have been worked out for a variety of different statistical  
5021 tests (Cohen 2013) and codified in software like G\*Power (Faul et al.  
5022 2007) and the pwr package for R (Champely 2020). For other cases  
5023 (including mixed effects models), you may have to conduct a simulation  
5024 in which you generate many simulated experimental runs under known  
5025 assumptions and compute how many of these lead to a significant effect;  
5026 luckily, R packages exist for this purpose as well, including the simr  
5027 package (Green and MacLeod 2016).

#### 5028 10.4.2 Power analysis in practice

5029 Let's do a power analysis for our hypothetical money and happiness  
5030 experiment. Imagine the experiment is a simple two-group design  
5031 in which participants from a convenience population are randomly  
5032 assigned either to receive \$1,000 and some advice on saving money  
5033 (experimental condition) vs just receiving the advice and no money  
5034 (control condition). We then follow up a month later and collect  
5035 self-reported happiness ratings. How many people should we have  
5036 in our study in order to be able to reject the null? The answer to

<sup>11</sup> Our focus here is on giving you a conceptual introduction to power analysis, but we refer you to Cohen (1992) for a more detailed introduction.

5037 this question depends on our desired values of  $\alpha$  and  $\beta$  as well as our  
5038 expected effect size for the intervention.

5039 For  $\alpha$  we will just set a conventional significance threshold of  $\alpha = 0.05$ .

5040 But what should be our desired level of power? The usual standard in  
5041 the social sciences is to aim for power above 80% (e.g.,  $\beta < 0.20$ ); this  
5042 gives you four out of five chances to observe a significant effect. But just  
5043 like  $\alpha = 0.05$ , this is a conventional value that is perhaps a little bit too  
5044 loose for modern standards—a strong test of a particular effect should  
5045 probably have 90% or 95% power.<sup>12</sup>

5046 These choices are relatively easy, compared to the fundamental issue:  
5047 our power analysis requires some expectation about our effect size. This  
5048 is the **first fundamental problem of power analysis**: if you knew the  
5049 effect size, you might not need to do the experiment!

5050 So how are you supposed to get an estimate of effect size? Here are a  
5051 few possibilities:

5052 – **Meta-analysis.** If there is a good meta-analysis of the effect that  
5053 you are trying to measure (or something closely related), then you  
5054 are in luck. A strong meta-analysis will have not only a precise ef-  
5055 fect size estimate but also some diagnostics detecting and correct-  
5056 ing potential publication bias in the literature (see chapter 16).

<sup>12</sup> Really, researchers interested in using power analysis in their work should give some thought to what sort of chance of a false negative they are willing to accept. In exploratory research, perhaps a higher chance of missing an effect is reasonable; in contrast, in confirmatory research it might make sense to aim for a higher level of power.

5057 While these diagnostics are imperfect, they still can give you a  
5058 sense for whether you can use the meta-analytic effect size esti-  
5059 mate as the basis for a power analysis.

5060 – **Specific prior study.** A more complicated scenario is when you  
5061 have only one or a handful of prior studies that you would like  
5062 to use as a guide. The trouble is that any individual effect in the  
5063 literature is likely to be inflated by publication and other selective  
5064 reporting biases (see chapter 3). Thus, using this estimate likely  
5065 means your study will be underpowered—you might not get as  
5066 lucky as a previous study did!

5067 – **Pilot testing.** Many people (including us) at some point learned  
5068 that one way to do a power analysis is to conduct a pilot study,  
5069 estimate the effect size from the pilot, and then use this effect  
5070 estimate for power analysis in the main study. We don't recom-  
5071 mend this practice. The trouble is that your pilot study will have  
5072 a small sample size, leading to a very imprecise estimate of effect  
5073 size (Browne 1995). If you overestimate the effect size, your main  
5074 study will be very underpowered. If you underestimate, the op-  
5075 posite will be true. Using a pilot for power analysis is a recipe for  
5076 problems.

5077 – **General expectations about an effect of interest.** In our view, per-

5078      haps the best way you can use power analysis (in the absence of a  
5079      really strong meta-analysis, at least) is to start with a general idea  
5080      about the size of effect you expect and would like to be able to  
5081      detect. It is totally reasonable to say, “I don’t know how big my ef-  
5082      fect is going to be, but let’s see what my power would be if it were  
5083      *medium-sized* (say  $d = 0.5$ ), since that’s the kind of thing we’re  
5084      hoping for with our money intervention.” This kind of power  
5085      analysis can help you set your expectations about what range of  
5086      effects you might be able to detect with a given sample size.

5087      For our money study, using our general expectation of a medium-size  
5088      effect, we can compute power for  $d = 0.5$ . In this case, we’ll simply use  
5089      the two-sample  $t$ -test introduced in chapter 6, for which 80% power at  
5090       $\alpha = 0.05$  and  $d = 0.5$  is achieved by having  $N = 64$  in each group.

### CODE

Classic power analysis in R is quite simple using the `pwr` package. The package offers a set of test-specific functions like `pwr.t.test()`. For each, you supply three of the four parameters specifying effect size (`d`), number of observations (`n`), significance level (`sig.level`), and power (`power`); the function computes the fourth. For classic power analysis, we leave out `n`:

```
pwr.t.test(d = .5,  
            power = .8,  
            sig.level = .05,  
            type = "two.sample",  
            alternative = "two.sided")
```

But it is also possible to use this same function to compute the power achieved at a combination of  $n$  and  $d$ , for example.

5092

5093 There's a second issue, however. The **second fundamental problem of**  
5094 **power analysis** is that the real effect size for an experiment may be zero.  
5095 And in that case, *no* sample size will let you correctly reject the null. Go-  
5096 ing back to our discussion in chapter 6, the null hypothesis significance  
5097 testing framework is just not set up to let you *accept* the null hypothesis.  
5098 If you are interested in a bidirectional approach to hypothesis testing  
5099 in which you can accept *and* reject the null, you may need to consider  
5100 Bayes Factor or **equivalence testing** approaches (Lakens, Scheel, and Is-  
5101 ager 2018), which don't fit the assumptions of classical power analysis.

#### 5102 10.4.3 Alternative approaches to sample size planning

5103 Let's now consider some alternatives to classic power analysis that can  
5104 still yield reasonable sample size justifications.

5105       1. **Resource constraint.** In some cases, there are fundamental re-  
5106       source constraints that limit data collection. For example, if you  
5107       are doing fieldwork, sometimes the right stopping criterion for  
5108       data collection is “when the field visit is over,” since every ad-  
5109       ditional datapoint is valuable. When prespecified, these kinds of  
5110       sample size justifications can be quite reasonable, although they  
5111       do not preclude being underpowered to test a particular hypoth-  
5112       esis.

5113       2. **Smallest effect size of interest (SESOI).** SESOI analysis is a variant  
5114       on power analysis that includes some resource constraint planning.  
5115       Instead of trying to intuit how big your target effect is, you instead  
5116       choose a level below which you might not be interested in detect-  
5117       ing the effect. This choice can be informed by theory (what is  
5118       predicted), applied concerns (what sort of effect might be useful  
5119       in a particular context), or resource constraints (how expensive or  
5120       time-consuming it might be to run an experiment). In practice,  
5121       SESOI analysis is simply a classic power analysis with a particular  
5122       small effect as the target.

5123       3. **Precision-based sample planning.** As we discussed in chapter 6,  
5124       the goal of research is not always to reject the null hypothesis!  
5125       Sometimes—we’d argue that it should be most of the time—the

goal is to estimate a particular causal effect of interest with a high level of precision, since these estimates are a prerequisite for building theories. If what you want is an estimate with known precision (say, a confidence interval of a particular width), you can compute the sample size necessary to achieve that precision (Bland 2009; Rothman and Greenland 2018).<sup>13</sup>

4. **Sequential analysis.** Your stopping rule need not be a hard cutoff at a specific  $N$ . Instead, it's possible to plan a **sequential analysis** using either frequentist or Bayesian methods, in which you plan to stop collecting data once a particular inferential threshold is reached. For the frequentist version, the key thing that keeps sequential analysis from being *p*-hacking is that you prespecify particular values of  $N$  at which you will conduct tests and then correct your *p*-values for having tested multiple times (Lakens 2014).

For Bayesian sequential analysis, you can actually compute a running Bayes factor as you collect data and stop when you reach a prespecified level of evidence (Schönbrodt et al. 2017). This latter alternative has the advantage of allowing you to collect evidence for the null as well as against it.<sup>14</sup>

In sum, there are many different ways of justifying your sample size or your stopping rule. The most important things are (1) to prespecify your

<sup>13</sup> In our experience, this kind of planning is most useful when you are attempting to gather measurements with sufficient precision to compare between computational models. Since the models can make quantitative predictions that differ by some known amount, then it's clear how tight your confidence intervals need to be.

<sup>14</sup> Another interesting variant is sequential parameter estimation, in which you collect data until a desired level of precision is achieved (Kelley, Darku, and Chattopadhyay 2018); this approach combines some of the benefits of both precision-based analysis and sequential analysis.

strategy and (2) to give a clear justification for your choice. Table 10.2 gives an example sample size justification that draws on several different concepts discussed here, using classical power computations as one part of the justification. A reviewer could easily follow the logic of this discussion and form their own conclusion about whether this study had an adequate sample size and whether it should have been conducted given the researchers' constraints.

Table 10.2  
Example sample size justification, referencing elements of SESOI, resource limitation, and power-based reasoning.

---

Element	Justification text
Background	We did not have strong prior information about the likely effect size, so we could not compute a classical power analysis.
Smallest effect of interest	Because of our interest in meaningful factors affecting word learning, we were interested in effect sizes as small as $d = 0.5$ .
Resource limitation	We were also limited by our ability to collect data only at our on-campus preschool.
Power computation	We calculated that based on our maximal possible sample size of $N = 120$ (60 per group), we would achieve at least 80% power to reject the null for effects as small as $d = 0.52$ .

---

 DEPTH

### *Sample sizes for replication studies*

Setting the sample size for a replication study has been a persistent issue in the metascience literature. Naïvely speaking, it seems like you should be able to compute the effect size for the original study and then simply use that as the basis for a classical power analysis.

This naive approach has several flaws, however. First, the effect size from the original published paper is likely an overestimate of the true effect size due to publication bias (Nosek et al. 2022). Second, the power analysis will only yield the sample size at which the replication will have a particular chance of rejecting the null at some criterion. But it's quite possible that the original experiment could be  $p < 0.05$ , the replication could be  $p > 0.05$ , *and* the original experiment and the replication results are not significantly different from each other. So a statistically significant replication of the original effect size is not necessarily what you want to aim for.

Faced with these issues, a replication sample size can be planned in several other ways. First, replicators can use the standard strategies mentioned above such as SESOI or resource-based planning to rule out large effects, either with high probability or within a known amount of time or money. If the SESOI is high or limited resources are allocated, these strategies can produce an inconclusive result, however. A conclusive answer can require a very substantial commitment of resources.

Second, Simonsohn (2015) recommends the “small telescopes” approach. The idea is not to test whether there *is* an effect, but rather where there is an effect *large enough that the original study could have detected it*. The analogy is to astronomy. If a birdwatcher points their binoculars at the sky and claims to have discovered a new planet, we want to ask not just whether there is a planet at that location but also whether there is any possibility that they could have seen it using binoculars—if not, perhaps they are right but for the wrong reasons! Simonsohn shows that, if a replicator collects 2.5 times as large a sample as the original, they have 80% power to detect any effect that was reasonably detectable by the original. This simple rule of thumb provides one good starting place for conservative replication studies.

Finally, replicators can make use of sequential Bayesian analysis, in which they attempt to gather substantial evidence relative to the support for  $H_1$  or  $H_0$ . Sequential bayes is an appealing option because it allows for efficient collection of data that reflects whether an effect is likely to be present in a particular sample, especially in the face of the sometimes prohibitively large samples necessary for SESOI or “small telescopes” analyses.

5155

## 5156 10.5 Chapter summary: Sampling

5157 Your goal as an experimenter is to estimate a causal effect. But the effect

5158 for whom? This chapter has tried to help you think about how you

5159 generalize from your experimental sample to some target population.

5160 It's very rare to be conducting an experiment based on a probability  
5161 sample in which every member of the population has an equal chance of  
5162 being selected. In the case that you are using a convenience sample, you  
5163 will need to consider how bias introduced by the sample could relate  
5164 to the effect estimate you observed. Do you think this effect is likely  
5165 to be very heterogeneous in the population? Are there theories that  
5166 suggest that it might be larger or smaller for the convenience sample  
5167 you recruited?

5168 Questions about generalizability and sampling depend on the precise  
5169 construct you are studying, and there is no mechanistic procedure for  
5170 answering them. Instead, you simply have to ask yourself: How does  
5171 my sampling procedure qualify the inference I want to make based on  
5172 my data? Being transparent about your reasoning can be very helpful—  
5173 both to you and to readers of your work who want to contextualize the  
5174 generality of your findings.



### DISCUSSION QUESTIONS

1. We want to understand human cognition generally, but do you think it is a more efficient research strategy to start by studying certain features of cognition (perception, for example) in WEIRD convenience populations and then later check our generalizations in non-WEIRD groups? What are the arguments against this efficiency-based strategy?

2. One alternative position regarding sampling is that the most influential experiments aren't generalizations of some number to a population; they are demonstration experiments that show that some particular effect is possible under some circumstances (think Milgram's conformity studies, see chapter 4). On this argument, the specifics of population sampling are often secondary. Do you think this position makes sense?
3. One line of argument says that we can't ever make generalizations about the human mind because so much of the historical human population is simply inaccessible to us (we can't do experiments on ancient Greek psychology). In other words, sampling from a particular population is *also* sampling a particular moment in time. How should we qualify our research interpretations to deal with this issue?

5176



## READINGS

- The original polemic article on the WEIRD problem: Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010). “The WEIRDest People in the World?” *Behavioral and Brain Sciences* 33 (2–3): 61–83.
- A very accessible introduction to power analysis from its originator: Cohen, Jacob (1992). “A Power Primer.” *Psychological Bulletin* 112 (1): 155.
- A thoughtful and in-depth discussion of generalizability issues: Yarkoni, Tal (2020). “The Generalizability Crisis.” *Behavioral and*

5177

*Brain Sciences* 45:1–37.

IV

5179

## EXECUTION

5180

<sup>5181</sup> *References*

- Bland, John Martin. 2009. “The Tyranny of Power: Is There a Better Way to Calculate Sample Size?” *British Medical Journal* 339 (7730): 1133–1135.
- Browne, Richard H. 1995. “On the Use of a Pilot Sample for Sample Size Determination.” *Statistics in Medicine* 14 (17): 1933–1940.
- Champely, Stephane. 2020. *pwr: Basic Functions for Power Analysis*. <https://CRAN.R-project.org/package=pwr>.
- Cohen, Jacob. 1992. “A Power Primer.” *Psychological Bulletin* 112 (1): 155–159.
- Cohen, Jacob. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.
- Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. 2018. “Generalizability of Heterogeneous Treatment Effect Estimates across Samples.” *Proceedings of the National Academy of Sciences* 115 (49): 12441–12446.
- Cunningham, Scott. 2021. *Causal Inference*. Yale University Press.
- DeJesus, Jasmine M., Maureen A. Callanan, Graciela Solis, and Susan A. Gelman. 2019. “Generic Language in Scientific Communication.” *Proceedings of the National Academy of Sciences* 116 (37): 18370–18377.
- Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. “G\* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences.” *Behavior Research Methods* 39 (2): 175–191.
- Green, Peter, and Catriona J. MacLeod. 2016. “simr: An R Package for Power Analysis of Generalised Linear Mixed Models by Simulation.” *Methods in Ecology and Evolution* 7 (4): 493–498. <https://doi.org/10.1111/2041-5182>

210X.12504.

Hedge, Craig, Georgina Powell, and Petroc Sumner. 2018. “The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences.” *Behavior Research Methods* 50 (3): 1166–1186.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?” *Behavioral and Brain Sciences* 33 (2-3): 61–83.

Kelley, Ken, Francis Bilson Darku, and Bhargab Chattopadhyay. 2018. “Accuracy in Parameter Estimation for a General Class of Effect Sizes: A Sequential Approach.” *Psychological Methods* 23 (2): 226–243.

Klein, Olivier, Tom E. Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C. Frank. 2018. “A Practical Guide for Transparency in Psychological Science.” *Collabra: Psychology* 4 (1): 20. <https://doi.org/10.1525/collabra.158>.

Lakens, Daniël. 2014. “Performing High-Powered Studies Efficiently with Sequential Analyses.” *European Journal of Social Psychology* 44 (7): 701–710.

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): 259–269. <https://doi.org/10.1177/2515245918770963>.

Little, Roderick J. A., and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.

Majid, Asifa, and Niclas Burenhult. 2014. “Odors Are Expressible in Language, as Long as You Speak the Right Language.” *Cognition* 130 (2): 266–270.

Majid, Asifa, and Nicole Kruspe. 2018. “Hunter-Gatherer Olfaction Is Special.” *Current Biology* 28 (3): 409–413.

Markus, Hazel R., and Shinobu Kitayama. 1991. “Culture and the Self: Implications for Cognition, Emotion, and Motivation.” *Psychological Review* 98 (2): 224–253.

McNemar, Quinn. 1946. “Opinion-Attitude Methodology.” *Psychological Bulletin* 43 (4): 289.

Neyman, Jerzy. 1992. “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection.” In *Breakthroughs in Statistics*, 123–150. Springer.

Nosek, Brian A., Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber Almenberg, Fiona Fidler, Joseph Hilgard, Melissa Kline, and Michèle B. Nuijten. 2022. “Replicability, Robustness, and Reproducibility in Psychological Science.” *Annual Review of Psychology* 73 (1): 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>.

Nunan, David, Jeffrey Aronson, and Clare Bankhead. 2018. “Catalogue of Bias: Attrition Bias.” *Evidence Based Medicine* 23 (1): 21–22.

Piantadosi, Steven T., and Edward Gibson. 2014. “Quantitative Standards for Absolute Linguistic Universals.” *Cognitive Science* 38 (4): 736–756.

Rohrer, Julia M. 2018. “Thinking Clearly about Correlations and Causation: Graphical Causal Models for Observational Data.” *Advances in Methods and Practices in Psychological Science* 1 (1): 27–42.

Rosenthal, Robert, and Ralph L. Rosnow. 1984. *Essentials of Behavioral Research: Methods and Data Analysis*. New York: McGraw-Hill.

- Rothman, Kenneth J., and Sander Greenland. 2018. “Planning Study Size Based on Precision Rather than Power.” *Epidemiology* 29 (5): 599–603.
- Schönbrodt, Felix D., Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. “Sequential Hypothesis Testing with Bayes Factors: Efficiently Testing Mean Differences.” *Psychol. Methods* 22 (2): 322–339.
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–1366. <https://doi.org/10.1177/0956797611417632>.
- Simonsohn, Uri. 2015. “Small Telescopes: Detectability and the Evaluation of Replication Results.” *Psychological Science* 26 (5): 559–569.
- Syed, Moin, and U. Kathawalla. 2020. “Cultural Psychology, Diversity, and Representation in Open Science.” In *Cultural Methods in Psychology: Describing and Transforming Cultures*, edited by Kate C McLean, 427–454. Oxford University Press.
- Tessler, Michael Henry, and Noah D. Goodman. 2019. “The Language of Generalization.” *Psychological Review* 126 (3): 395–436.
- Tsai, Jeanne L. 2007. “Ideal Affect: Cultural Causes and Behavioral Consequences.” *Perspectives on Psychological Science* 2 (3): 242–259.
- Westfall, Jacob, Charles M. Judd, and David A. Kenny. 2015. “Replicating Studies in Which Samples of Participants Respond to Samples of Stimuli.” *Perspectives on Psychological Science* 10 (3): 390–399.
- Yarkoni, Tal. 2020. “The Generalizability Crisis.” *Behavioral and Brain Sciences* 45:1–37.

Robert Crosnoe, Chandra Muller, Elizabeth Tipton, et al. 2019. "A National Experiment Reveals Where a Growth Mindset Improves Achievement." *Nature* 573 (7774): 364–369.

Zhou, Haotian, and Ayelet Fishbach. 2016. "The Pitfall of Experimenting on the Web: How Unattended Selective Attrition Leads to Surprising (Yet False) Research Conclusions." *Journal of Personality and Social Psychology* 111 (4): 493–504.

## 5187 11 PREREGISTRATION

### apple LEARNING GOALS

- Recognize the dangers of researcher degrees of freedom
- Understand the differences between exploratory and confirmatory modes of research
- Articulate how preregistration can reduce risk of bias and increase transparency

5188

5189 When not planned beforehand, data analysis can approxi-

5190

mate a projective technique, such as the Rorschach, because

5191

the investigator can project on the data his own expectan-

5192

cies, desires, or biases and can pull out of the data almost

5193

any “finding” he may desire.

5194

—Theodore X. Barber (1976 [1927])

5195

The first principle is that you must not fool yourself—and

5196 you are the easiest person to fool ... After you've not fooled  
5197 yourself, it's easy not to fool other scientists. You just have  
5198 to be honest in a conventional way after that.

5199 —Richard Feynman (1974)

5200 The last section of the book focused on planning a study—in particular,  
5201 making decisions around measurement, design, and sampling. In this  
5202 next section, we turn to the nuts and bolts of executing a study. We  
5203 start with preregistration (this chapter), before discussing the logistics  
5204 of data collection (chapter 12) and project management (chapter 13).  
5205 These chapters touch on the themes of TRANSPARENCY and BIAS REDUC-  
5206 TION through decisions about how to document and organize your data  
5207 collection.

5208 Let's start with simply documenting choices about design and analysis.  
5209 Although there are plenty of *incorrect* ways to design and analyse exper-  
5210 iments, there is no single *correct* way. In fact, most research decisions  
5211 have many justifiable choices—sometimes called “researcher degrees of  
5212 freedom.” For example, will you stop data collection after 20, 200, or  
5213 2,000 participants? Will you remove outlier values, and how will you  
5214 define them? Will you conduct subgroup analyses to see whether the  
5215 results are affected by sex, or age, or some other factor?

5216 Consider a simplified, hypothetical case where you have to make five  
5217 analysis decisions and there are five justifiable choices for each decision—  
5218 this alone would result in 3,125 ( $5^5$ ) unique ways to analyze the data!  
5219 If you were to make these decisions post hoc (after observing the data)  
5220 then there's a danger your decisions will be influenced by the outcome  
5221 of the analysis ("data-dependent decision making") and skew toward  
5222 choices that generate outcomes more aligned with your personal pref-  
5223 erences. Now think back to the last time you read a research paper. Of  
5224 all the possible ways that the data could have been analyzed, how do you  
5225 know that the researchers did not just select the approach that generated  
5226 results most favorable to their pet hypothesis?

5227 In this chapter, we will find out why flexibility in the design, analy-  
5228 sis, reporting, and interpretation of experiments, combined with data-  
5229 dependent decision-making, can introduce bias and lead to scientists  
5230 fooling themselves and each other. We will also learn about **preregistra-**  
5231 **tion**, the process of writing down and registering your research decisions  
5232 before you observe the data. Preregistration intersects with two of our  
5233 themes: it can be used to REDUCE BIAS in our data analysis, and it can pro-  
5234 vide the TRANSPARENCY that other scientists need to properly evaluate  
5235 and interpret our results (Hardwicke and Wagenmakers 2023).

 CASE STUDY

## *Undisclosed analytic flexibility?*

Educational apps for children are a huge market, but relatively few randomized trials have been done to see whether or when they produce educational gains. Filling this important gap, Berkowitz et al. (2015) reported a high-quality field experiment of a free educational app, “Bedtime Math at Home,” with participants randomly assigned to either math or reading conditions over the course of a full school year. Critically, along with random assignment, the study also included standardized measures of math and reading achievement. These measures allowed the authors to compute effects in grade-level equivalents, a meaningful unit from a policy perspective.

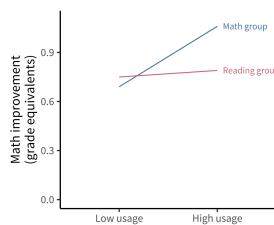


Figure 11.1

Model fits reported in figure 1 of Berkowitz et al. (2015). Estimated years of math achievement gained over the school year across groups, as a function of app usage level.

The key result is shown in figure 11.1. Families who used the math app frequently showed greater gains in math than the control group. Although this finding appeared striking, the figure didn't directly visualize the primary causal effect of interest, namely the size of the effect of study condition on math scores. Instead the data were presented as estimated effects

for specific levels of app usage.

Because the authors made their data openly available, it was possible for Frank (2016) to do a simple analysis to examine the causal effect of interest. When not splitting the data by usage and adjusting by covariates, there was no significant main effect of the intervention on math performance figure 11.2. Since this analysis was not favorable to the primary intervention—and because it was not reported in the paper—it could have been the case that the authors had analyzed the data several ways and chosen to present an analysis that was more favorable to their hypotheses of interest.

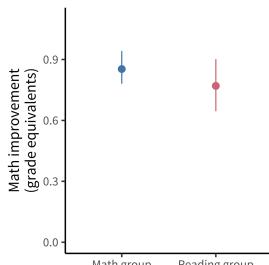


Figure 11.2  
Estimated years of math achievement gained over the school year across groups in the Berkowitz et al. (2016) math app trial. Error bars show bootstrapped 95% confidence intervals. Based on Frank (2016).

As is true for many papers prior to the rise of preregistration, it's not possible to know definitively whether the reported analysis in Berkowitz et al. (2015) was influenced by the authors' desired result. As we'll see below, such data-dependent analyses can lead to substantial bias in reported effects. This uncertainty about a paper's analytic strategy can be avoided by the use of preregistration. In this case, preregistration would have convinced readers that the analyses decisions were not influenced by the data,

thereby increasing the value of this otherwise high-quality study.

5238

### 5239 11.1 *Lost in a garden of forking paths*

5240 One way to visualize researcher degrees of freedom is as a vast decision  
5241 tree or “garden of forking paths” (figure 11.3). Each node represents  
5242 a decision point, and each branch represents a justifiable choice. Each  
5243 unique pathway through the garden terminates in an individual research  
5244 outcome.

5245 Because scientific observations typically consist of both noise (random  
 5246 variation unique to this sample) and signal (regularities that will reoccur  
 5247 in other samples), some of these pathways will inevitably lead to results  
 5248 that are misleading (e.g., inflated effect sizes, exaggerated evidence, or  
 5249 false positives). The more potential paths in the garden that you might  
 5250 explore, the higher the chance of encountering misleading results.

5251 Statisticians refer to this issue as a **multiplicity** (multiple comparisons)  
 5252 problem. As we talked about in chapter 6, multiplicity can be addressed  
 5253 to some extent with statistical countermeasures, like the Bonferroni cor-  
 5254 rection; however, these adjustment methods need to account for every  
 5255 path that you *could have* taken (Gelman and Loken 2014; de Groot 2014  
 5256 [1956]). When you navigate the garden of forking paths while work-  
 5257 ing with the data, it is easy to forget—or even be unaware of—every  
 5258 path that you could have taken, so these methods can no longer be used  
 5259 effectively.

5260 The signal-to-noise ratio is worse in particular situations (as com-  
 5261 mon in psychology) with small effect sizes, high variation, and large  
 5262 measurement errors (Ioannidis 2005). Researcher degrees of freedom  
 5263 may be constrained to some extent by strong theory (Oberauer and  
 5264 Lewandowsky 2019), community methodological norms, or replication  
 5265 studies, though these constraints may be more implicit than explicit,

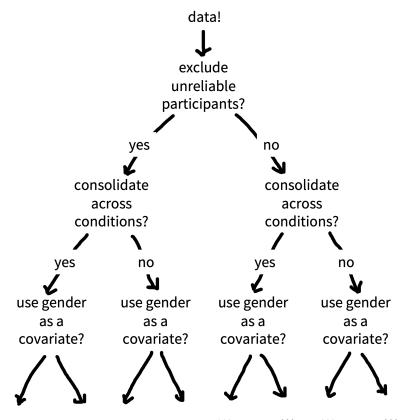


Figure 11.3

The garden of forking paths: many justifiable but different analytic choices are possible for an individual dataset.

5266 and can still leave plenty of room for flexible decision-making.

5267 *11.1.1 Data-dependent analysis*

5268 When a researcher navigates the garden of forking paths during data  
5269 analysis, their choices might be influenced by the data (**data-dependent**  
5270 **decision-making**), which can introduce bias. If a researcher is seeking a  
5271 particular kind of result (see the DEPTH box below), then they are more  
5272 likely to follow the branches that steer them in that direction.

5273 You could think of this a bit like playing a game of “hot (🔥) or cold  
5274 (❄️),” where 🔥 indicates that the choice will move the researcher closer  
5275 to a desirable overall result and ❄️ indicates that the choice will move  
5276 them further away. Each time the researcher reaches a decision point,  
5277 they try one of the branches and get feedback on how that choice af-  
5278 fects the results. If the feedback is 🔥 then they take that branch. If the  
5279 answer is ❄️, they try a different branch. If they reach the end of a com-  
5280 plete pathway, and the result is ❄️, maybe they even retrace their steps  
5281 and try some different branches earlier in the pathway. This strategy  
5282 creates a risk of bias because it systematically skews results toward the  
5283 researcher’s preferences (Hardwicke and Wagenmakers 2023).<sup>1</sup>

<sup>1</sup> We say “risk of bias” rather than just “bias” because in most scientific contexts, we do not have a known ground truth to compare the results to. So in any specific situation, we do not know the extent to which data-dependent decisions have actually biased the results.

 DEPTH

## *Only human: Cognitive biases and skewed incentives*

There's a storybook image of the scientist as an objective, rational, and dispassionate arbiter of truth (Veldkamp et al. 2017). But in reality, scientists are only human: they have egos, career ambitions, and rent to pay! So even if we do want to live up to the storybook image, it's important to acknowledge that our decisions and behavior are also influenced by a range of cognitive biases and external incentives that can steer us away from that goal. Let's first look at some relevant cognitive biases that might lead scientists astray:

- **Confirmation bias:** Preferentially seeking out, recalling, or evaluating information in a manner that reinforces one's existing beliefs (Nickerson 1998).
- **Hindsight bias:** Believing that past events were always more likely to occur relative to our actual belief in their likelihood before they happened ("I knew it all along!") (Slovic and Fischhoff 1977).
- **Motivated reasoning:** Rationalizing prior decisions so they are framed in a favorable light, even if they were irrational (Kunda 1990).
- **Apophenia:** Detecting seemingly meaningful patterns in noise (Gilovich, Vallone, and Tversky 1985).

To make matters worse, the incentive structure of the scientific ecosystem often adds additional motivation to get things wrong. The allocation

of funding, awards, and publication prestige is often based on the nature of research results rather than research quality (Smaldino and McElreath 2016; Nosek, Spies, and Motyl 2012). For example, many academic journals, especially those that are widely considered to be the most prestigious, appear to have a preference for novel, positive, and statistically significant results over incremental, negative, or null results (Bakker, Dijk, and Wicherts 2012). There is also pressure to write articles with concise, coherent, and compelling narratives (Giner-Sorolla 2012). This set of forces incentivizes scientists to be “impressive” over being right and encourages questionable research practices. The process of iteratively *p*-hacking and HARKing one’s way to a “beautiful” scientific paper has been dubbed “The Chrysalis Effect” (O’Boyle, Banks, and Gonzalez-Mulé 2017), illustrated in figure 11.4.

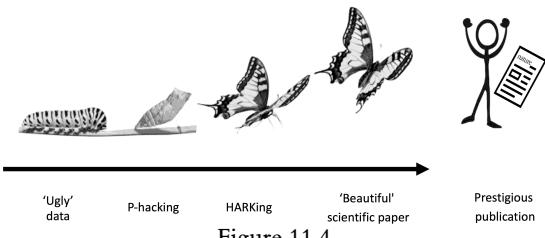


Figure 11.4  
The Chrysalis Effect, when ugly truth becomes a beautiful fiction.

In sum, scientists’ human flaws—and the scientific ecosystem’s flawed incentives—highlight the need for transparency and intellectual humility when reporting the findings of our research (Hoekstra and Vazire 2021).

5286 In the most egregious cases, a researcher may try multiple pathways un-  
5287 til they obtain a desirable result and then **selectively report** that result,  
5288 neglecting to mention that they have tried several other analysis strate-  
5289 gies (also known as *p*-hacking, a practice we've discussed throughout  
5290 the book).<sup>2</sup> You may remember an example of this practice in chap-  
5291 ter 3, where participants apparently became younger when they listened  
5292 to "When I'm 64" by The Beatles. Another example of how dam-  
5293 aging the garden of forking paths can be comes from the "discovery"  
5294 of brain activity in a dead Atlantic Salmon (Bennett, Miller, and Wol-  
5295 ford 2009)! Researchers deliberately exploited flexibility in the fMRI  
5296 analysis pipeline and avoided multiple comparisons corrections, allow-  
5297 ing them to find brain activity where there was only dead fish (fig-  
5298 ure 11.5).

### 5299 11.1.1 Hypothesizing after results are known

5300 In addition to degrees of freedom in experimental design and analysis,  
5301 there is additional flexibility in how researchers *interpret* research results.  
5302 As we discussed in chapter 2, theories can accommodate even conflict-  
5303 ing results in many different ways—for example, by positing auxiliary  
5304 hypotheses that explain why a particular datapoint is special.  
5305 The practice of selecting or developing your hypothesis after observing

<sup>2</sup> "If you torture the data long enough, it will confess" (Good 1972).

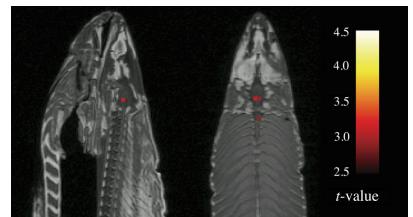
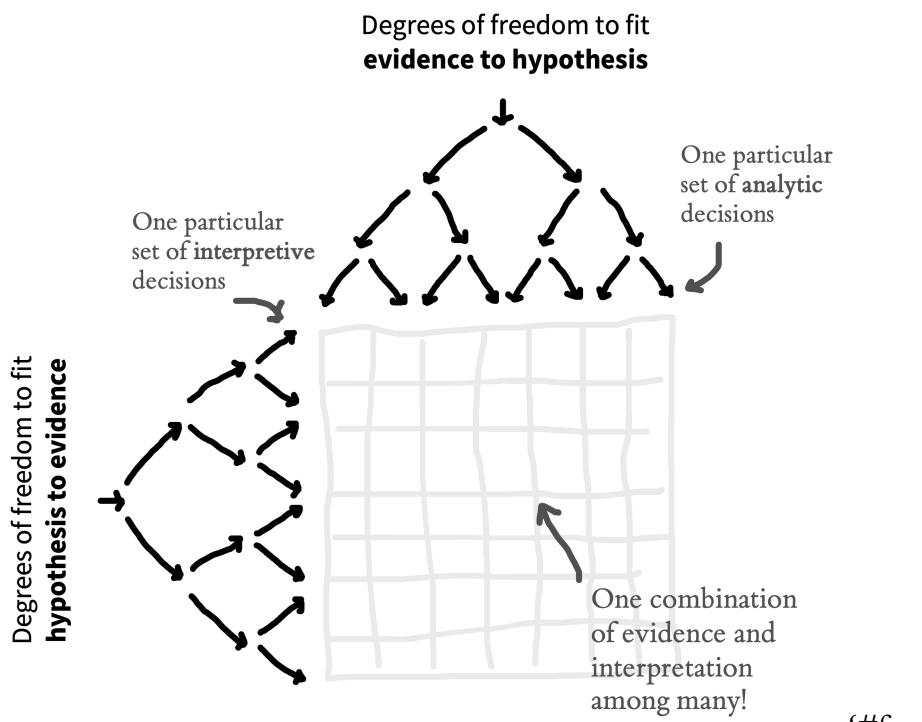


Figure 11.5  
By deliberately exploiting analytic flexi-  
bility in the processing pipeline of fMRI  
data, Bennett, Miller, and Wolford  
(2009) were able to identify "brain ac-  
tivity" in a dead Atlantic Salmon. From  
Bennett, Miller, and Wolford (2009, li-  
censed under CC BY).

5306 the data has been called “hypothesizing after the results are known,” or  
 5307 “HARKing” (Kerr 1998). HARKing is potentially problematic because  
 5308 it expands the garden of forking paths and helps to justify the use of var-  
 5309 ious additional design and analysis decisions (?@fig-grid). For example,  
 5310 you may come up with an explanation for why an intervention is effec-  
 5311 tive in men but not in women in order to justify a post hoc subgroup  
 5312 analysis based on sex (see the CASE STUDY box). The extent to which  
 5313 HARKing is problematic is contested (for discussion see Hardwicke and  
 5314 Wagenmakers 2023). But at the very least, it’s important to be honest  
 5315 about whether hypotheses were developed before or after observing the  
 5316 data.



5318 grid .margin-caption width=70% fig-alt="A diagram of a grid with axes  
5319 degrees of freedom to fit" evidence to hypotheses" and "hypotheses to  
5320 evidence".")

5321 But hang on a minute! Isn't it a good thing to seek out interesting results  
5322 if they are there in the data? Shouldn't we "let the data speak"? The  
5323 answer is yes! But it's crucial to understand the distinction between **ex-**  
5324 **ploratory and confirmatory** modes of research.<sup>3</sup> Confirmation involves  
5325 making research decisions *before* you've seen the data whereas explo-  
5326 ration involves making research decisions *after* you've seen data.

<sup>3</sup> In practice, an individual study may contain both exploratory and confirmatory aspects, which is why we describe them as different "modes."

5327 The key things to remember about exploratory research are that you  
5328 need to (1) be aware of the increased risk of bias arising from data-  
5329 dependent decision-making and calibrate your confidence in the results  
5330 accordingly; and (2) be honest with other researchers about your anal-  
5331 ysis strategy so they are also aware of the risk of bias and can calibrate  
5332 *their* confidence in the outcomes accordingly. In the next section, we  
5333 will learn about how preregistration helps us to make this important  
5334 distinction between exploratory and confirmation research.

5335 11.2 Reducing risk of bias, increasing transparency, and

5336 calibrating confidence with preregistration

5337 You can counter the problem of researcher degrees of freedom and data-

5338 dependent decision-making by making research decisions before you

5339 have seen the data—like planning your route through the garden of fork-

5340 ing paths before you start your journey (Wagenmakers et al. 2012; Hard-

5341 wicke and Wagenmakers 2023). If you stick to the planned route, then

5342 you have eliminated the possibility that your decisions were influenced

5343 by the data.

5344 Preregistration is the process of declaring your research decisions in

5345 a public registry before you analyze (and often before you collect)

5346 the data. Preregistration ensures that your research decisions are

5347 data-independent, which reduces risk of bias arising from the issues

5348 described above. Preregistration also transparently conveys to others

5349 what you planned, helping them to determine the risk of bias and

5350 calibrate their confidence in the research results. In other words,

5351 preregistration can dissuade researchers from engaging in questionable

5352 research practices like *p*-hacking and HARKing, because they can be

5353 held accountable to their original plan while also providing the context

5354 needed to properly evaluate and interpret research.

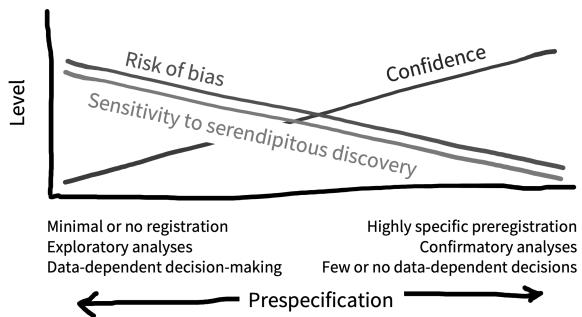


Figure 11.6

Preregistration clarifies where research activities fall on the continuum of pre-specification. When the preregistration provides little constraint over researcher degrees of freedom (i.e., more exploratory research), decisions are more likely to be data-dependent, and consequently there is a higher risk of bias. When preregistration provides strong constraint over researcher degrees of freedom (i.e., more confirmatory research), decisions are less likely to be data dependent, and consequently there is a lower risk of bias. Exploratory research activities are more sensitive to serendipitous discovery but also have a higher risk of bias relative to confirmatory research activities. Preregistration transparently communicates where particular results are located along the continuum, helping readers to appropriately calibrate their confidence.

- 5355 Preregistration does not require that you specify all research decisions  
 5356 in advance, only that you are transparent about what was planned, and  
 5357 what was not planned. This transparency helps to make a distinction be-  
 5358 tween which aspects of the research were exploratory and which were  
 5359 confirmatory (figure 11.6). All else being equal, we should have more  
 5360 confidence in confirmatory results, because there is a lower risk of bias.  
 5361 Exploratory results have a higher risk of bias, but they are also more  
 5362 sensitive to serendipitous (unexpected) discoveries. So the confirma-  
 5363 tory mode is best suited to testing hypotheses, and the exploratory mode  
 5364 is best suited to generating them. Therefore, exploratory and confirma-  
 5365 tory research are both valuable activities—it is just important to differen-  
 5366 tiate them (Tukey 1980)! Preregistration offers the best of both worlds  
 5367 by clearly separating one from the other.  
 5368 In addition to the benefits described above, preregistration may improve  
 5369 the quality of research by encouraging closer attention to study plan-  
 5370 ning. We've found that the process of writing a preregistration really

5371 helps facilitate communication between collaborators, and can catch ad-

5372 dressable problems before time and resources are wasted on a poorly de-

5373 signed study. Detailed advanced planning can also create opportunities

5374 for useful community feedback, particularly in the context of registered

5375 reports (see the DEPTH box below), where dedicated peer reviewers will

5376 evaluate your study before it has even begun.

 DEPTH

## *Preregistration and friends: A toolbox to address researcher degrees of freedom*

Several useful tools can be used to complement or extend preregistration.

In general, we would recommend that these tool are combined with pre-registration, rather than used as a replacement because preregistration provides transparency about the research and planning process (Hardwicke and Wagenmakers 2023). The first two of these are discussed in more detail in the last section of chapter 7.

**Robustness checks.** Robustness checks (also called “sensitivity analyses”) assess how different decision choices in the garden of forking paths affect the eventual pattern of results. This technique is particularly helpful when you have to choose between several justifiable analytic choices, neither of which seem superior to the other, or which have complementary strengths and weaknesses. For example, you might run the analysis three times using three different methods for handling missing data. Robust results should not vary substantially across the three different choices.

**Multiverse analyses.** Recently, some researchers have started running large-scale robustness checks called “multiverse” (Steegen et al. 2016) or “specification curve” (Simonsohn, Simmons, and Nelson 2020) analyses. We discussed these a bit in chapter 7. Some have argued that these large-scale robustness checks make preregistration redundant; after all, why pre-

specify a single path if you can explore them all (Rubin 2020; Oberauer and Lewandowsky 2019)? But interpreting the results of a multiverse analysis is not straightforward; for example, it seems unlikely that all of the decision choices are equally justifiable (Giudice and Gangestad 2021). Furthermore, if multiverse analyses are not preregistered, then they introduce researcher degrees of freedom and create an opportunity for selective reporting, which increases risk of bias.

**Held-out sample.** One option to benefit from both exploratory and confirmatory research modes is to split your data into **training** and **test** samples. (The test sample is commonly called “held out” because it is “held out” from the exploratory process.) You can generate hypotheses in an exploratory mode in the training sample and use that as the basis to preregister confirmatory analyses in the held-out sample. A notable disadvantage of this strategy is that splitting the data reduces statistical power, but in cases where data are plentiful—including in much of machine learning—this technique is the gold standard.

**Masked analysis** (traditionally called “blind analysis”). Sometimes problems, such as missing data, attrition, or randomization failure that you did not anticipate in your preregistered plan, can arise during data collection. How do you diagnose and address these issues without increasing risk of bias through data-dependent analysis? One option is masked analysis, which disguises key aspects of the data related to the results (for example, by shuffling condition labels or adding noise) while still allowing some degree of data inspection (Dutilh, Sarafoglou, and Wagenmakers 2019). Af-

ter diagnosing a problem, you can adjust your preregistered plan without increasing risk of bias, because your decisions have not been influenced by the results.

**Standard operating procedures.** Community norms, perhaps at the level of your research field or lab, can act as a natural constraint on researcher degrees of freedom. For example, there may be a generally accepted approach for handling outliers in your community. You can make these constraints explicit by writing them down in a standard operating procedures (SOP) document—a bit like a living meta-preregistration (Lin and Green 2016).

**Open lab notebooks.** Maintaining a lab notebook can be a useful way to keep a record of your decisions as a research project unfolds. Preregistration is a bit like taking a snapshot of your lab notebook at the start of the project, when all you have written down is your research plan. Making your lab notebook publicly available is a great way to transparently document your research and departures from the preregistered plan.



Figure 11.7

Registered reports (from <https://www.cos.io/initiatives/registered-reports>, licensed under CC BY 4.0).

**Registered reports.** Registered reports (figure 11.7) are a type of article format that embeds preregistration directly into the publication pipeline (Chambers and Tzavella 2020). The idea is that you submit your prereg-

istered protocol to a journal and it is peer reviewed before you've even started your study. If the study is approved, the journal agrees to publish it, regardless of the results. This is a radical departure from traditional publication models where peer reviewers and journals evaluate your study *after* its been completed and the results are known. Because the study is accepted for publication independently of the results, registered reports can offer the benefits of preregistration with additional protection against publication bias. They also provide a great opportunity to obtain feedback on your study design while you can still change it!

5380

### 5381 11.3 How to preregister

5382 High-stakes studies such as medical trials must be preregistered (Dick-  
5383 ersin and Rennie 2012). In 2005, a large international consortium of  
5384 medical journals decided that they would not publish unregistered trials.  
5385 The discipline of economics also has strong norms about study registra-  
5386 tion (see, e.g., <https://www.socialscienceregistry.org>). But preregistra-  
5387 tion is pretty new to psychology (Nosek et al. 2018), and there's still no  
5388 standard way of doing it—you're already at the cutting edge!

5389 We recommend using the Open Science Framework (OSF) as your reg-  
5390 istry. OSF is one of the most popular registries in psychology, and you  
5391 can do lots of other useful things on the platform to make your research

5392 transparent, like sharing data, materials, analysis scripts, and preprints.

5393 On OSF, it's possible to "register" any file you have uploaded. When

5394 you register a file, it creates a time-stamped, read-only copy, with a ded-

5395 icated link. You can add this link to articles reporting your research.

Table 11.1  
Preregistration template outline.

Question
1 Data collection. Have any data been collected for this study already?
2 Hypothesis. What's the main question being asked or hypothesis being tested in this study?
3 Dependent variable. Describe the key dependent variable(s) specifying how they will be measured.
4 Conditions. How many and which conditions will participants be assigned to?
5 Analyses. Specify exactly which analyses you will conduct to examine the main question/hypothesis.
6 Outliers and Exclusions. Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
7 Sample Size. How many observations will be collected, or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
8 Other. Anything else you would like to preregister (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned).

5396 One approach to preregistration is to write a protocol document that  
 5397 specifies the study rationale, aims or hypotheses, methods, and analysis  
 5398 plan, and register that document.<sup>4</sup> Open Science Framework also has  
 5399 a collection of dedicated preregistration templates that you can use if  
 5400 you prefer. An outline of such a template is shown in table 11.1. These  
 5401 templates are often tailored to the needs of particular types of research.

<sup>4</sup> You can think of a study protocol as a bit like a research paper without a results and discussion section (here's an example from one of our own studies: <https://osf.io/2cnkq>).

5402 For example, there are templates for general quantitative psychology  
5403 research (“PRP-QUANT”; Bosnjak et al. 2022), cognitive modeling  
5404 (Crüwell and Evans 2021), and secondary data analysis (Akker et al.  
5405 2019). The OSF interface may change, but currently this guide<sup>5</sup>  
5406 provides a set of steps to create a preregistration.

5407 Once you’ve preregistered your plan, you just go off and run the study  
5408 and report the results, right? Well hopefully … but things might not  
5409 turn out to be that straightforward. It’s quite common to forget to in-  
5410 clude something in your plan or to have to depart from the plan due  
5411 to something unexpected. Preregistration can actually be pretty hard in  
5412 practice (Nosek et al. 2019).

5413 Don’t worry though—remember that a key goal of preregistration  
5414 is transparency to enable others to evaluate and interpret research  
5415 results. If you decide to depart from your original plan and conduct  
5416 data-dependent analyses, then this decision may increase the risk of  
5417 bias. But if you communicate this decision transparently to your  
5418 readers, they can appropriately calibrate their confidence in the results.

5419 You may even be able to run both the planned and unplanned analyses  
5420 as a robustness check (see the DEPTH box) to evaluate the extent to  
5421 which this particular choice impacts the results.

5422 When you report your study, it is important to distinguish between

<sup>5</sup> <https://help.osf.io/article/158-create-a-preregistration>

5423 what was planned and what was not. If you ran a lot of data-dependent  
5424 analyses, then it might be worth having separate exploratory and confir-  
5425 matory results sections. On the other hand, if you mainly stuck to your  
5426 original plan, with only minor departures, then you could include a ta-  
5427 ble (perhaps in an appendix) that outlines these changes (for example,  
5428 see Supplementary Information A of this article<sup>6</sup>).

<sup>6</sup> [https://doi.org/10.31222/osf.io/  
wt5ny](https://doi.org/10.31222/osf.io/wt5ny)

#### 5429 11.4 Chapter summary: Preregistration

5430 We've advocated here for preregistering your study plan. This practice  
5431 helps to reduce the risk of bias caused by data-dependent analysis (the  
5432 "garden of forking paths" that we described) and transparently commu-  
5433 nicate the risk of bias to other scientists. Importantly, preregistration is  
5434 a "plan, not a prison"<sup>7</sup>: in most cases, preregistered, confirmatory anal-  
5435 yses coexist with exploratory analyses. Both are an important part of  
5436 good research—the key is to disclose which is which!

<sup>7</sup> [https://www.cos.io/blog/  
preregistration-plan-not-prison](https://www.cos.io/blog/preregistration-plan-not-prison)



#### DISCUSSION QUESTIONS

1. P-hack your way to scientific glory! To get a feel for how data-dependent analyses might work in practice, have a play around with this app: <https://projects.fivethirtyeight.com/p-hacking>. Do you think preregistration would affect your confidence in claims made about this dataset?

2. Preregister your next experiment! The best way to get started with preregistration is to have a go with your next study. Head over to <https://osf.io/registries/osf/new> and register your study protocol or complete one of the templates. What aspects of preregistration did you find most difficult, and what benefits did it bring?

5438

## READINGS

- Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor (2018). “The Preregistration Revolution.” *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606. <https://doi.org/10.1073/pnas.1708274114>.
- Hardwicke, Tom E., and Eric-Jan Wagenmakers (2023). “Reducing Bias, Increasing Transparency, and Calibrating Confidence with Pre-registration.” *Nature Human Behaviour* 7 (1): 15–26. <https://doi.org/10.31222/osf.io/d7bcu>.

5439

## 5440 *References*

- Akker, Olmo van den, Sara J. Weston, Lorne Campbell, William J. Chopik, Rodica I. Damian, Pamela Davis-Kean, Andrew Hall, et al. 2019. “Pre-registration of Secondary Data Analysis: A Template and Tutorial.” PsyArXiv. <https://psyarxiv.com/hvfmr/>.
- Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts. 2012. “The Rules of the Game Called Psychological Science.” *Perspectives on Psychological*

5441

- Science 7 (6): 543–554. <https://doi.org/10.1177/1745691612459060>.
- Barber, Theodore Xenophon. 1976 [1927]. *Pitfalls in Human Research: Ten Pivotal Points*. Pergamon General Psychology Series. Pergamon Press.
- Bennett, C. M., M. B. Miller, and G. L. Wolford. 2009. “Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument for Multiple Comparisons Correction.” *NeuroImage*, Organization for Human Brain Mapping 2009 Annual Meeting, 47 (Supplement 1): S125. [https://doi.org/10.1016/S1053-8119\(09\)71202-9](https://doi.org/10.1016/S1053-8119(09)71202-9).
- Berkowitz, Talia, Marjorie W. Schaeffer, Erin A. Maloney, Lori Peterson, Courtney Gregor, Susan C. Levine, and Sian L. Beilock. 2015. “Math at Home Adds up to Achievement in School.” *Science* 350 (6257): 196–198. <https://doi.org/10.1126/science.aac7427>.
- Berkowitz, Talia, Marjorie W. Schaeffer, Christopher S. Rozek, Erin A. Maloney, Susan C. Levine, and Sian L. Beilock. 2016. “Response to Comment on ‘Math at Home Adds up to Achievement in School’.” *Science* 351 (6278): 1161.
- Bosnjak, Michael, Christian Fiebach, David Thomas Mellor, Stefanie Mueller, Daryl O’Connor, Fred Oswald, and Rose Sokol-Chang. 2022. “A Template for Preregistration of Quantitative Research in Psychology: Report of the Joint Psychological Societies Preregistration Task Force.” *American Psychologist* 77 (4): 602–615. <https://doi.org/10.1037/amp0000879>.
- Chambers, Chris, and Loukia Tzavella. 2020. “Registered Reports: Past, Present and Future.” MetaArXiv. <https://doi.org/10.31222/osf.io/43298>.
- Crüwell, Sophia, and Nathan J. Evans. 2021. “Preregistration in Diverse Con-

texts: A Preregistration Template for the Application of Cognitive Models.” *Royal Society Open Science* 8 (10): 210155. <https://doi.org/10.1098/rsos.210155>.

de Groot, A. D. 2014 [1956]. “The Meaning of ‘Significance’ for Different Types of Research.” Translated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier A. Kievit, Marjan Bakker, Angélique O. J. Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica* 148 (May 2014): 188–194. <https://doi.org/10.1016/j.actpsy.2014.02.001>.

Dickersin, Kay, and Drummond Rennie. 2012. “The Evolution of Trial Registries and Their Use to Assess the Clinical Trial Enterprise.” *JAMA* 307 (17): 1861–1864. <https://doi.org/10.1001/jama.2012.4230>.

Dutilh, Gilles, Alexandra Sarafoglou, and Eric-Jan Wagenmakers. 2019. “Flexible Yet Fair: Blinding Analyses in Experimental Psychology.” *Synthese*, August. <https://doi.org/10.1007/s11229-019-02456-7>.

Feynman, Richard P. 1974. “Cargo Cult Science.” <http://caltech.library.caltech.edu/51/2/CargoCult.pdf>.

Frank, Michael C. 2016. “Comment on ‘Math at Home Adds up to Achievement in School’.” *Science* 351 (6278): 1161.

Gelman, Andrew, and Eric Loken. 2014. “The Statistical Crisis in Science.” *American Scientist* 102 (6): 460–465. <https://doi.org/10.1511/2014.111.460>.

Gilovich, Thomas, Robert Vallone, and Amos Tversky. 1985. “The Hot Hand in Basketball: On the Misperception of Random Sequences.” *Cognitive Psychology* 17 (3): 295–314. [https://doi.org/10.1016/0010-0285\(85\)5443](https://doi.org/10.1016/0010-0285(85)5443)

90010–6.

Giner-Sorolla, Roger. 2012. “Science or Art? How Aesthetic Standards Grease the Way through the Publication Bottleneck but Undermine Science.” *Perspectives on Psychological Science* 7 (6): 562–571. <https://doi.org/10.1177/1745691612457576>.

Giudice, M. Del, and S. W. Gangestad. 2021. “A Traveler’s Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions.” *Advances in Methods and Practices in Psychological Science* 4 (1): 1–15. <https://doi.org/10.1177/2515245920954925>.

Good, I. J. 1972. “Statistics and Today’s Problems.” *The American Statistician* 26 (3): 11–19. <https://doi.org/10.1080/00031305.1972.10478922>.

Hardwicke, Tom E., and Eric-Jan Wagenmakers. 2023. “Reducing Bias, Increasing Transparency, and Calibrating Confidence with Preregistration.” *Nature Human Behaviour* 7 (1): 15–26. <https://doi.org/10.1038/s41562-022-01497-2>.

Hoekstra, Rink, and Simine Vazire. 2021. “Aspiring to Greater Intellectual Humility in Science.” *Nature Human Behaviour* 5 (12): 1602–1607.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLOS Medicine* 2 (8): e124. <https://doi.org/10.1371/journal.pmed.0020124>.

Kerr, Norbert L. 1998. “HARKing: Hypothesizing after the Results Are Known.” *Personality & Social Psychology Review (Lawrence Erlbaum Associates)* 2 (3): 196. [https://doi.org/10.1207/s15327957pspr0203\\_4](https://doi.org/10.1207/s15327957pspr0203_4).

Kunda, Ziva. 1990. “The Case for Motivated Reasoning.” *Psychological Bulletin* 108 (3): 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>.

Lin, Winston, and Donald P. Green. 2016. “Standard Operating Procedures: A Safety Net for Pre-Analysis Plans.” *PS: Political Science & Politics* 49 (3): 495–500. [http://www.journals.cambridge.org/abstract\\_S1049096516000810](http://www.journals.cambridge.org/abstract_S1049096516000810).

Nickerson, Raymond S. 1998. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” *Review of General Psychology* 2 (2): 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>.

Nosek, Brian A., Emorie D. Beck, Lorne Campbell, Jessica K. Flake, Tom E. Hardwicke, David T. Mellor, Anna E. van ’t Veer, and Simine Vazire. 2019. “Preregistration Is Hard, and Worthwhile.” *Trends in Cognitive Sciences* 23 (10): 815–818. <https://doi.org/10.1016/j.tics.2019.07.009>.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. “The Preregistration Revolution.” *Proceedings of the National Academy of Sciences* 115 (11): 2600–2606. <https://doi.org/10.1073/pnas.1708274114>.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. “Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability.” *Perspectives on Psychological Science* 7 (6): 615–631. <https://doi.org/10.1177/1745691612459058>.

O’Boyle, Ernest Hugh, George Christopher Banks, and Erik Gonzalez-Mulé. 2017. “The Chrysalis Effect: How Ugly Initial Results Metamorphosize into Beautiful Articles.” *Journal of Management* 43 (2): 376–399. <https://doi.org/10.1177/0149206314527133>.

Oberauer, Klaus, and Stephan Lewandowsky. 2019. “Addressing the Theory Crisis in Psychology.” *Psychonomic Bulletin & Review* 26 (5): 1596–1618.

Rubin, Mark. 2020. “Does Preregistration Improve the Credibility of Research Findings?” *The Quantitative Methods for Psychology* 16 (4): 15. <https://doi.org/10.20982/tqmp.16.4.p376>.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. “Specification Curve Analysis.” *Nature Human Behaviour*, July, 1–7. <https://doi.org/10.1038/s41562-020-0912-z>.

Slovic, Paul, and Baruch Fischhoff. 1977. “On the Psychology of Experimental Surprises.” *Journal of Experimental Psychology: Human Perception and Performance* 3 (4): 544–551. <https://doi.org/10.1037/0096-1523.3.4.544>.

Smaldino, Paul E., and Richard McElreath. 2016. “The Natural Selection of Bad Science.” *Royal Society Open Science* 3 (9): 160384. <https://doi.org/10.1098/rsos.160384>.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. “Increasing Transparency through a Multiverse Analysis.” *Perspectives on Psychological Science* 11 (5): 702–712. <https://doi.org/10.1177/1745691616658637>.

Tukey, John W. 1980. “We Need Both Exploratory and Confirmatory.” *The American Statistician* 34 (1): 23–25. <https://doi.org/10.2307/2682991>.

Veldkamp, Coosje L. S., Chris H. J. Hartgerink, Marcel A. L. M. van van Assen, and Jelte M. Wicherts. 2017. “Who Believes in the Storybook Image of the Scientist?” *Accountability in Research* 24 (3): 127–151. <https://doi.org/10.1080/08989621.2016.1268922>.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit. 2012. “An Agenda for Purely Confirmatory Research.” *Perspectives on Psychological Science* 7 (6): 632–638. <https://doi.org/10.1177/1745691612454912>.

5447      org/10.1177/1745691612463078.

5448

## 12 DATA COLLECTION



### LEARNING GOALS

- Outline key features of informed consent and participant debriefing
- Identify additional protections necessary for working with vulnerable populations
- Review best practices for online and in-person data collection
- Implement data integrity checks, manipulation checks, and pilot testing

5449

5450 You have selected your measure and manipulation and planned your  
5451 sample. Your preregistration is set. Now it's time to think about the  
5452 nuts and bolts of collecting data. Though the details may vary between  
5453 contexts, this chapter will describe some general best practices for data  
5454 collection.<sup>1</sup> We organize our discussion of these practices around two  
5455 perspectives: the participant and the researcher.

5456 The first section takes the perspective of a participant. We begin by

<sup>1</sup> The metaphor of “collection” implies to some researchers that the data exist independent of the researcher’s own perspective and actions, so they reject it in favor of the term “data generation.” Unfortunately, this alternative label doesn’t distinguish generating data via interactions with participants on the one hand and generating data from scratch via statistical simulations on the other. We worry that “data generation” sounds too much like the kinds of fraudulent data generation that we talked about in chapter 4, so we have opted to keep the more conventional “data collection” label.

5457 reviewing the importance of informed consent. A key principle of run-  
5458 ning experiments with human participants is that we respect their au-  
5459 tonomy, which includes their right to understand the study and choose  
5460 whether to take part. When we neglect the impact of our research on  
5461 the people we study, we not only violate regulations governing research;  
5462 we also create distrust that undermines the moral basis of scientific re-  
5463 search.

5464 In the second section, we begin to shift perspectives, discussing the  
5465 choice of online vs in-person data collection and some of the advan-  
5466 tages of online data collection for TRANSPARENCY. We consider how to  
5467 optimize the experimental experience for participants in both settings.  
5468 We then end by taking the experimenter's perspective more fully, dis-  
5469 cussing how we can maximize data quality (contributing to MEASURE-  
5470 MENT PRECISION) using pilot testing, manipulation checks, and attention  
5471 checks, while still being cognizant of both changes to the participant's  
5472 experience and the integrity of statistical inferences (both contributing  
5473 to BIAS REDUCTION).

### CASE STUDY

#### *The rise of online data collection*

Since the rise of experimental psychology laboratories in university set-  
tings during the period after World War II (Benjamin 2000), experiments

have typically been conducted by recruiting participants from what has been referred to as the “subject pool.” This term denotes a group of people who can be recruited for experiments, typically students from introductory psychology courses (Sieber and Saks 1989) who are required to complete a certain number of experiments as part of their course work. The ready availability of this convenient population inevitably led to the massive overrepresentation of undergraduates in published psychology research, undermining its generalizability (Sears 1986; Henrich, Heine, and Norenzayan 2010).

Yet, over the last couple of decades, there has been a revolution in data collection. Instead of focusing on university undergraduates, increasingly researchers recruit individuals from crowdsourcing websites like Amazon Mechanical Turk and Prolific Academic. Crowdsourcing services were originally designed to recruit and pay workers for ad hoc business tasks like retyping receipts, but they have also become marketplaces to connect researchers with research participants who are willing to complete surveys and experimental tasks for small payments (Litman, Robinson, and Abberbock 2017). As of 2015, more than a third of studies in top social and personality psychology journals were conducted on crowdsourcing platforms (another third were still conducted with college undergraduates), and this proportion is likely continuing to grow (Anderson et al. 2019).

Initially, many researchers worried that crowdsourced data from online convenience samples would lead to a decrease in data quality. However,

several studies suggest that data quality from online convenience samples is typically comparable to in-lab convenience samples (Mason and Suri 2012; M. Buhrmester, Kwang, and Gosling 2011). In one particularly compelling demonstration, a set of online experiments were used to replicate a group of classic phenomena in cognitive psychology, with clear successes on every experiment except those requiring sub-50 millisecond stimulus presentation (Crump, McDonnell, and Gureckis 2013). Further, as we discuss below, researchers have developed a suite of tools to ensure that online participants understand and comply with the instructions in complex experimental tasks.

Since these initial successes, however, attention has moved away from the validity of online experiments to the ethical challenges of engaging with crowdworkers. In 2020, nearly 130,000 people completed MTurk studies (Moss et al. 2020). Of those, an estimated 70% identified as White, 56% identified as women, and 48% had an annual household income below \$50,000. A sampling of crowd work determined that the average wage earned was just \$2.00 per hour, and less than 5% of workers were paid at least the federal minimum wage (Hara et al. 2018). Further, many experimenters routinely withheld payment from workers based on their performance in experiments. These practices clearly violate ethical guidelines for research with human participants but are often overlooked by institutional review boards who may be unfamiliar with online recruitment platforms or consider that platforms are offering a “service” rather than simply being alternative routes for paying individuals.

With greater attention to the conditions of workers (e.g., Salehi et al. 2015), best practices for online research have progressed considerably. As we describe below, working with online populations requires attention to both standard ethical issues of consent and compensation, as well as new issues around the “user experience” of participating in research. The availability of online convenience samples can be transformative for the pace of research, for example, by enabling large studies to be run in a single day rather than over many months. But online participants are vulnerable in different ways than university convenience samples, and we must take care to ensure that research online is conducted ethically.

5477

### 5478 *12.1 Informed consent and debriefing*

5479 As we discussed in chapter 4, experimenters must respect the autonomy  
5480 of their participants: they must be informed about the risks and bene-  
5481 fits of participation before they agree to participate. Researchers must  
5482 also discuss and contextualize the research by debriefing participants af-  
5483 ter they have completed the study. Here we look at the nuts and bolts  
5484 of each of these processes, ending with guidance on the special protec-  
5485 tions that are required to protect the autonomy of especially vulnerable  
5486 populations.

5487 12.1.1 *Getting consent*

5488 Experimental participants must give consent. In most regulatory frame-  
5489 works, there are clear guidelines about what the process of giving con-  
5490 sent should look like. Typically participants are expected to read and  
5491 sign a **consent form**: a document that explains the goals of the research  
5492 and its procedures, describes potential risks and benefits, and asks for  
5493 participants' explicit consent to participate voluntarily. Table 12.1 gives  
5494 the full list of consent form requirements from the US Office for Hu-  
5495 man Research Protections, and figure 12.1 shows how these individual  
5496 requirements are reflected in a real consent form used in our research.

Table 12.1  
US Office of Human Research Protections requirements for a consent form (edited for length).

Requirement
1 A statement that the study involves research
2 An explanation of the purposes of the research
3 The expected duration of the subject's participation
4 A description of the procedures to be followed
5 Identification of any procedures that are experimental
6 A description of any reasonably foreseeable risks or discomforts to the subject
7 A description of any benefits to the subject or to others that may reasonably be expected from the research
8 A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the subject

---

Requirement

---

- 9 A statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained
  - 10 For research involving more than minimal risk, an explanation as to whether any compensation or medical treatments are available if injury occurs
  - 11 An explanation of whom to contact for answers to pertinent questions about the research and research subjects' rights
  - 12 A statement that participation is voluntary, refusal to participate will involve no penalty, and that subject may discontinue participation at any time without penalty
-

<b>STANFORD UNIVERSITY Research Consent Form</b>		<b>IRB USE ONLY</b>
Protocol Director: Michael C. Frank, Ph.D.		Approval Date:
Protocol Title: Investigations of language learning and social cognition in infants, children and adults		Expiration Date:
<p><b>DESCRIPTION:</b> In this study, we are investigating the development of language and communication. Our research explores how infants and young children learn about their native language. We observe how children at different ages learn new words and comprehend familiar words. All of the activities in our studies are designed to be age-appropriate and fun for children. In a typical session, we may invite your child to play a short game, or we may use an eye-tracker (a special camera that keeps track of where a child is looking on a computer screen) to help us understand what your child is looking at while they listen to recorded speech and/or view movies of adults, children, puppets, or animated characters playing and talking. Sometimes some of the speech they hear will be from a foreign or made-up language.</p>		
<p><b>2 RISKS AND BENEFITS:</b> There are no foreseeable risks or discomforts to you or your child in participating in this research. All our procedures are observational and non-intrusive. We pace each session appropriately and give breaks as needed to enable your child to enjoy and complete the session. Your child will not be pressured to continue in the event that he or she becomes upset, tired, or resistant at any point during the session. If for any reason you or your child do not want to continue, the session will be ended immediately with no penalty.</p>		
<p><b>7</b> We cannot and do not guarantee or promise that you will receive any benefits from this study, apart from the honorarium and the satisfaction of participating in developmental research. If appropriate, we provide information regarding resources that may be helpful in addressing any concerns regarding your child's development.</p>		
<p><b>3 TIME INVOLVEMENT:</b> Each session typically lasts from 5-10 minutes, depending on the nature of the study. Most studies involve a single session, but in some cases you and your child will be invited to participate in more than one session.</p>		
<p><b>6 PAYMENTS:</b> You will not receive a cash payment for your participation in this research. However, based on the number and length of sessions we arrange with you during scheduling, your child will receive one of the following gifts in appreciation of your time and cooperation: a children's book, T-shirt, or certificate of appreciation.</p>		
<p><b>12 SUBJECT'S RIGHTS:</b> If you have read this form and have decided to allow your child to participate in this project, please understand your child's participation is voluntary and your child has the right to withdraw his/her consent or discontinue participation at any time without penalty or loss of benefits to which he/she is otherwise entitled. Your child has the right to refuse to answer particular questions. The video record of the session will be identified by a code number, not by name. This record will be accessible only to the project director and members of the project staff, unless you give your explicit permission below for others to view it for scientific or educational purposes. All records will be stored securely so that your child's individual privacy will be maintained. In addition, your child's identity will remain private in all publications resulting from the study.</p>		
<p><b>11 CONTACT INFORMATION:</b></p> <ul style="list-style-type: none"> <li>* Questions, Concerns, or Complaints: If you have any questions, concerns or complaints about this research study, its procedures, risks, and benefits you should contact the Protocol Director, Dr. Michael Frank, phone: (650) 721-9270, email: <a href="mailto:langcoglab@stanford.edu">langcoglab@stanford.edu</a>, webpage: <a href="http://langcog.stanford.edu">http://langcog.stanford.edu</a></li> <li>* Independent Contact: If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact the Stanford Institutional Review Board (IRB) to speak to someone independent of the research team at (650) 723-2480 or toll free at 1-866-680-2906. You can also write to the Stanford IRB, Stanford University, 1705 El Camino Real, Palo Alto, CA 94306.</li> </ul>		
<p><b>CONSENT</b> I give consent for my child to be videotaped during this study. please initial: _____ Yes _____ No</p>		
<p>I give consent for your child's image (from the video recording) to be shown to people not directly involved with this research during class, seminars, or scientific presentations. please initial: _____ Yes _____ No</p>		
<p>Please sign below.</p>		
<p>Signature of Parent, Guardian or Conservator _____ Date _____  <small>The IRB determined that the permission of one parent is sufficient for research to be conducted under 45 CFR 46.404, in accordance with 45 CFR 46.408(b).</small></p>		
<p>The extra copy of this consent form is for you to keep.</p>		
<p>For Office Use Only   Study: _____ SubjID: _____</p>		

Figure 12.1

A consent form annotated to show how specific text fulfills the requirements in table 12.1. Categories 5, 8, and 10 were not required for this minimal-risk psychology experiment.

- 5497 These are just samples. Since ethics regulation is almost always man-  
 5498 aged at the institutional level, your local ethics board will often provide  
 5499 guidance on the specific information you should include in the consent  
 5500 form and they will almost always need to approve the form before you  
 5501 are allowed to begin recruiting participants.
- 5502 When providing consent information, researchers should focus on what

5503 someone might think or feel as a result of participating in the study. Are  
5504 there any physical or emotional risks associated? What should someone  
5505 know about the study that may give them pause about agreeing to par-  
5506 ticipate in the first place? Our advice is to center the *participant* in the  
5507 consent process rather than the research question. Information about  
5508 specific research goals can typically be provided during debriefing.<sup>2</sup>

5509 If there are specific pieces of information about study goals or proce-  
5510 dures that *must* be withheld from participants during consent, **decep-**  
5511 **tion** of participants may be warranted. Deception can be approved by  
5512 ethics boards as long as it poses little risk and is effectively addressed via  
5513 more extensive debriefing. But an experimental protocol that includes  
5514 deception will likely undergo greater scrutiny during ethics review, as  
5515 it must be justified by a specific experimental need.

5516 During the consent process, researchers should explain to participants  
5517 what will be done with their data. Requirement 9 in table 12.1 asks for  
5518 a statement about data confidentiality, but such a statement is a mere  
5519 minimum. Some modern consent forms explicitly describe different  
5520 uses of the data and ask for consent for each. For example, the form in  
5521 figure 12.1 asks permission for showing recordings as part of presenta-  
5522 tions.<sup>3</sup>

<sup>2</sup> Some experimenters worry that informing participants about the study that they are about to participate in may influence their behavior in the study via so-called demand characteristics, discussed in chapter 9. But the goal of a consent form is not to explain the specific psychological construct being manipulated. Instead, a consent form typically focuses on the experience of being in the study (for example, that a participant would be asked to provide quick verbal responses to pictures). This sort of general explanation should not create demand characteristics.

5523    *12.1.2 Prerequisites of consent*

5524    To give consent, participants must have the cognitive capacity to make  
5525    decisions (competence), understand what they are being asked to do  
5526    (comprehension), and know that they have the right to withdraw con-  
5527    sent at any time (voluntariness) (Kadam 2017).

5528    Typically, we assume competence for adult volunteers in our experi-  
5529    ments, but if we are working with children or other vulnerable popula-  
5530    tions (see below), we may need to consider whether they are legally com-  
5531    petent to provide consent. Participants who cannot consent on their  
5532    own should still be informed about participation in an experiment, and,  
5533    if possible, you should still obtain their **assent** (informal agreement) to  
5534    participate. When a person has no legal ability to consent, you must  
5535    obtain consent from their legal guardian. But if they do not assent, you  
5536    should also respect their decision not to participate—even if you previ-  
5537    ously obtained consent from their guardian.

5538    The second prerequisite is comprehension. It is good practice to dis-  
5539    cuss consent forms verbally with participants, especially if the study is  
5540    involved and takes place in person. If the study is online, ensure that  
5541    participants know how to contact you if they have questions about the  
5542    study. The consent form itself must be readable for a broad audience,

<sup>3</sup> Some ethics boards will ask for con-  
sent for sharing even anonymized data  
files. As we discuss in chapter 13, fully  
anonymized data can often be shared  
without explicit consent. You may still  
choose to ask participants' permission,  
but this practice may lead to an awkward  
situation—for example, a dataset with  
heterogeneous sharing permissions such  
that most but not all data can be shared  
publicly. Norms around anonymized  
data sharing are shifting, so it's worth  
having a conversation with your ethics  
board about how they interpret your par-  
ticular regulatory obligations.

meaning care should be taken to use accessible language and clear formatting. Consider giving participants a copy of the consent form in advance so they can read at their own pace, think of any questions they might have, and decide how to proceed without any chance of feeling coerced (Young, Hooker, and Freeberg 1990).

Finally, participants must understand that their involvement is voluntary, meaning that they are under no obligation to be involved in a study and always have the right to withdraw at any time. Experimenters should not only state that participation is voluntary; they should also pay attention to other features of the study environment that might lead to structural coercion (Fisher 2013). For example, high levels of compensation can make it difficult for lower-income participants to withdraw from research. Similarly, factors like race, gender, and social class can lead participants to feel discomfort around discontinuing a study. It is incumbent on experimenters to provide a comfortable study environment and to avoid such coercive factors wherever possible.

### 5559 12.1.3 Debriefing participants

Once a study is completed, researchers should always debrief participants. A debriefing is composed of four parts: (1) gratitude, (2) discussion of goals, (3) explanation of deception (if relevant), and (4) questions

5563 and clarification (Allen 2017). Together these serve to contextualize the  
5564 experience for the participant and to mitigate any potential harms from  
5565 the study.

5566 **1. Gratitude.** Thank participants for their contribution! Sometimes  
5567 thanks is enough (for a short experiment), but many studies also  
5568 include monetary compensation or course credit. Compensation  
5569 should be commensurate with the amount of time and effort re-  
5570 quired for participation. Compensation structures vary widely  
5571 from place to place; typically local ethics boards will have specific  
5572 guidelines.

5573 **2. Discussion of goals.** Researchers should share the purpose of the  
5574 research with participants in, aiming for a short and accessible  
5575 statement that avoids technical jargon. Sharing goals is especially  
5576 important when some aspect of the study appears evaluative—  
5577 participants will often be interested in knowing how well they  
5578 performed against their peers. For example, a parent whose child  
5579 completed a word-recognition task may request information  
5580 about their child’s performance. It can assuage parents’ worries  
5581 to highlight that the goals of the study are about measuring a  
5582 particular experimental effect, not about individual evaluation  
5583 and ranking.<sup>4</sup>

<sup>4</sup> At the study’s conclusion, you might also consider sharing any findings with participants—many participants appreciate learning about research findings that they contributed to, even months or years after participation.

5584        3. **Explanation of deception.** Researchers must reveal any deception  
5585                  during debriefing, regardless of how minor the deception seems  
5586                  to the researcher. This component of the debriefing process can  
5587                  be thought of as “dehoaxing” because it is meant to illuminate  
5588                  any aspects of the study that were previously misleading or inac-  
5589                  curate (Holmes 1976). The goal is both to reveal the true intent  
5590                  of the study and to alleviate any potential anxiety associated with  
5591                  the deception. Experimenters should make clear both where in  
5592                  the study the deception occurred and why the deception was nec-  
5593                  essary for the study’s success.

5594        4. **Questions and clarification.** Finally, researchers should answer  
5595                  any questions or address any concerns raised by participants.  
5596                  Many researchers use this opportunity to ask participants about  
5597                  their own ideas about the study goals. This practice not only  
5598                  illuminates aspects of the study design that may have been unclear  
5599                  to or hidden from participants, but also begins a discussion where  
5600                  both researchers and participants can communicate about this  
5601                  joint experience. This step is also helpful in identifying negative  
5602                  emotions or feelings resulting from the study (Allen 2017).  
5603                  When participants do express negative emotions, researchers are  
5604                  responsible for sharing resources participants can use to help  
5605                  them.<sup>5</sup>

<sup>5</sup> In the case that participants report substantial concerns or negative reactions to an experiment—during debriefing or otherwise—researchers will typically have an obligation to report these to their ethics board.

5606    *12.1.4 Special considerations for vulnerable populations*

5607    Regardless of who is participating in research, investigators have an obli-  
5608    gation to protect the rights and well-being of all participants. Some pop-  
5609    ulations are considered especially **vulnerable** because of their decreased  
5610    agency—either in general or in the face of potentially coercive situa-  
5611    tions. Research with these populations receives additional regulatory  
5612    oversight. In this section, we will consider several vulnerable popula-  
5613    tions.

5614    **Children.** Children are some of the most commonly used vulner-  
5615    able populations in research because the study of development can  
5616    contribute both to children’s welfare and to our understanding of  
5617    the human mind. In the US, children under the age of 18 may only  
5618    participate in research with written consent from a parent or guardian.  
5619    Unless they are preverbal, children should additionally be asked for  
5620    their assent. The risks associated with a research study focusing on  
5621    children also must be no greater than minimal unless participants may  
5622    receive some direct benefit from participating or participating in the  
5623    study may improve a disorder or condition the participant was formally  
5624    diagnosed with.

5625    **People with disabilities.** There are thousands of disabilities that affect  
5626    cognition, development, motor ability, communication, and decision-

5627 making with varying degrees of interference, so it is first important to  
5628 remember that considerations for this population will be just as diverse  
5629 as its members. No laws preclude people with disabilities from partici-  
5630 pating in research. However, those with cognitive disabilities who are  
5631 unable to make their own decisions may only participant with written  
5632 consent from a legal guardian and with their individual assent (if ap-  
5633 plicable). Those retaining full cognitive capacity but who have other  
5634 disabilities that make it challenging to participate normally in the study  
5635 should receive appropriate assistance to access information about the  
5636 study, including the risks and benefits of participation.

5637 **Incarcerated populations.** Nearly 2.1 million people are incarcerated in  
5638 the United States alone (Gramlich 2021). Due to early (and repugnant)  
5639 use of prisoners as a convenience population that could not provide con-  
5640 sent, the use of prisoners in research has been a key focus of protective  
5641 efforts. The US Office for Human Research Protections (OHRP) sup-  
5642 ports their involvement in research under very limited circumstances—  
5643 typically when the research specifically focuses on issues relevant to in-  
5644 carcerated populations (Office for Human Research Protections 2003).

5645 When researchers propose to study incarcerated individuals, the local  
5646 ethics board must reconfigure to include at least one active prisoner (or  
5647 someone who can speak from a prisoner's perspective) and ensure that  
5648 less than half of the board has any affiliation to the prison system, pub-

lic or private. Importantly, researchers must not suggest or promise that participation will have any bearing on an individual's prison sentence or parole eligibility, and compensation must be otherwise commensurate with their contribution.

**Low-income populations.** Participants with fewer resources may be more persuaded to participate by monetary incentives, creating a potentially coercive situation. Researchers should consult with their local ethics board to conform to local standards for noncoercive payment.

**Indigenous populations.** There is a long and negative history of the involvement of indigenous populations in research without their consent. In the case that research requires the participation of indigenous individuals—because of potential benefits to their communities, rather than due to convenience—then community leadership must be involved to discuss the appropriateness of the research as well as how the consent process should be structured (Fitzpatrick et al. 2016).

**Crowdworkers.** Ethics boards do not usually consider crowdworkers on platforms like Amazon Mechanical Turk to be a specific vulnerable population, but many of the same concerns about diminished autonomy and greater need for protection still arise (see the Depth Box below). Without platform or ethics board standards, it is up to individual experimenters to commit to fair pay, which should ideally match or exceed

5670 the applicable minimum wage (e.g., the US federal minimum wage).  
5671 Further, in the context of reputation management systems like those of  
5672 Amazon Mechanical Turk, participants can be penalized for withdraw-  
5673 ing from an experiment—once they have their work “rejected” by an  
5674 experimenter, it can be harder for them to find new jobs, causing serious  
5675 long-term harm to their ability to earn on the platform.

## 5676 *12.2 Designing the “research experience”*

5677 For the majority of psychology experiments, the biggest factor that gov-  
5678 erns whether a participant has a positive or negative experience of an ex-  
5679 periment is not its risk profile, since for many psychology experiments  
5680 the quantifiable risk to participants is minimal.<sup>6</sup> Instead, it is the partic-  
5681 ipants’ experience. Did they feel welcome? Did they understand the  
5682 instructions? Did the software work as designed? Was their compensa-  
5683 tion clearly described and promptly delivered? These aspects of “user  
5684 experience” are critical both for ensuring that participants have a good  
5685 experience in the study (an ethical imperative) and for gathering good  
5686 data. An experiment that leaves participants unhappy typically doesn’t  
5687 satisfy either the ethical or the scientific goals of research. In this sec-  
5688 tion, we’ll discuss how to optimize the research experience for both  
5689 in-person and online experiments, as well as providing some guidance

<sup>6</sup> There are of course exceptions, includ-  
ing research with more sensitive content.  
Even in these cases, however, attention  
to the participant’s experience can be  
important for ensuring good scientific  
outcomes.

5690 on how to decide between these two administration contexts.

5691 *12.2.1 Ensuring good experiences for in-lab participants*

5692 A participant's experience begins even before they arrive at the lab.

5693 Negative experiences with the recruitment process (e.g., unclear con-

5694 sent forms, poor communication, or complicated scheduling) or tran-

5695 sit to the lab (e.g., difficulty navigating or finding parking) can lead to

5696 frustrated participants with a negative view of your research. Anything

5697 you can do to make these experiences smoother and more predictable—

5698 prompt communication, well-tested directions, reserved parking slots,

5699 and so on—will make your participants happier and increase the quality

5700 of your data.<sup>7</sup>

5701 Once a participant enters the lab, every aspect of the interaction with

5702 the experimenter can have an effect on their measured behavior (Gass

5703 and Seiter 2018). For example, a likable and authoritative experimenter

5704 who clearly describes the benefits of participation is following general

5705 principles for persuasion (Cialdini and Goldstein 2004). This interac-

5706 tion should lead to better compliance with experimental instructions,

5707 and hence better data, than an interaction with an unclear or indiffer-

5708 ent experimenter.

<sup>7</sup> For some reason, the Stanford Psychology Department building is notoriously difficult to navigate. This seemingly minor issue has resulted in a substantial number of late, frustrated, and flustered participants over the years.

5709 Any interaction with participants must be scripted and standardized so  
5710 that all participants have as similar an experience as possible. A lack  
5711 of standardization can result in differential treatment for participants  
5712 with different characteristics, which could result in data with greater  
5713 variability or even specific sociodemographic biases. An experimenter  
5714 that was kinder and more welcoming to one demographic group would  
5715 be acting unethically, and they also might find a very different result  
5716 than they intended.

5717 Even more importantly, experimenters who interact with participants  
5718 should ideally be unaware of the experimental condition each partic-  
5719 ipant is assigned to. This practice is often called “blinding” or “mask-  
5720 ing.” Otherwise it is easy for experimenter knowledge to result in small  
5721 differences in interaction across conditions, which in turn can influence  
5722 participants’ behavior, resulting in experimenter expectancy effects (see  
5723 chapter 9). Even if the experimenter must know a participant’s condi-  
5724 tion assignment—as is sometimes the case—this information should be  
5725 revealed at the last possible moment to avoid contamination of other  
5726 aspects of the experimental session.<sup>8</sup>

<sup>8</sup> In some experiments, an experimenter delivers a manipulation and, hence, it cannot be masked from them. In such cases, it’s common to have two experimenters such that one delivers the manipulation and another (masked to condition) collects the measurements. This situation often comes up with studies of infancy, since stimuli are often delivered via an in-person puppet show; at a minimum, behavior should be coded by someone other than the puppeteer.

5727 *12.2.2 Ensuring good experiences for online participants*

5728 The design challenges for online experiments are very different than for  
5729 in-lab experiments. As the experimental procedure is delivered through  
5730 a web browser, experimenter variability and potential expectancy ef-  
5731 fects are almost completely eliminated. On the other hand, some online  
5732 participants do many hours of online tasks a day and many are multi-  
5733 tasking in other windows or on other devices. It can be much harder  
5734 to induce interest and engagement in your research when your manip-  
5735 ulation is one of dozens the participant has experienced that day and  
5736 when your interactions are mediated by a small window on a computer  
5737 screen.

5738 When creating an online experimental experience, we consider four  
5739 issues: (1) design, (2) communication, (3) payment policies, and (4) ef-  
5740 fective consent and debriefing.<sup>9</sup>

<sup>9</sup> For extensive further guidance on this topic, see Litman and Robinson (2020).

5741 **Basic UX design.** Good experiment design online is a subset of good  
5742 web user experience (UX) design more generally. If your experi-  
5743 ment is unpleasant to interact with, participants will likely become con-  
5744 fused and frustrated. They will either drop out or provide data that are  
5745 lower quality. A good interface should be clean and well-tested and  
5746 should offer clear places where the participant must type or click to

5747 interact. If a participant presses a key at an appropriate time, the experi-  
5748 ment should offer a response—otherwise the participant will likely press  
5749 it again. If the participant is uncertain how many trials are left, they  
5750 may be more likely to drop out of the experiment so it is also helpful  
5751 to provide an indication of their progress. And if they are performing  
5752 a speeded paradigm, they should receive practice trials to ensure that  
5753 they understand the experiment prior to beginning the critical blocks  
5754 of trials.

5755 **Communication.** Many online studies involve almost no direct contact  
5756 with participants. When participants do communicate with you, it is  
5757 very important to be responsive and polite (as it is with in-lab partic-  
5758 ipants, of course). Unlike the typical undergraduate participant, the  
5759 work that a crowdworker is doing for your study may be part of how  
5760 they earn their livelihood, and a small issue in the study for you may  
5761 feel very important for them. For that reason, rapid resolution of is-  
5762 sues with studies—typically through appropriate compensation—is very  
5763 important. Crowdworkers often track the reputation of specific labs  
5764 and experimenters (sometimes through forums or specialized software;  
5765 [Irani and Silberman 2013](#)). A quick and generous response to an issue  
5766 will ensure that future crowdworkers do not avoid your studies.

5767 **Payment policies.** Unclear or punitive payment policies can have a ma-

5768 jor impact on crowdworkers. We strongly recommend *always* paying  
5769 workers if they complete your experiment, regardless of result. This  
5770 policy is comparable to standard payment policies for in-lab work. We  
5771 assume good faith in our participants: if someone comes to the lab, they  
5772 are paid for the experiment, even if it turns out that they did not per-  
5773 form correctly. The major counterargument to this policy is that some  
5774 online marketplaces have a population of workers who are looking to  
5775 cheat by being noncompliant with the experiment (e.g., entering gibber-  
5776 ish or even using scripts or artificial intelligence tools to progress quickly  
5777 through studies). Our recommendation is to address this issue through  
5778 the thoughtful use of “check” trials (see below)—not through punitive  
5779 nonpayment. The easiest way for a participant to complete your exper-  
5780 iment should be by complying with your instructions.

Table 12.2  
Sample online consent statement from our course.

---

By answering the following questions, you are participating in a study being performed by cognitive scientists in the Stanford Department of Psychology. If you have questions about this research, please contact us at stanfordpsych251@gmail.com. You must be at least 18 years old to participate. Your participation in this research is voluntary. You may decline to answer any or all of the following questions. You may decline further participation, at any time, without adverse consequences. Your anonymity is assured; the researchers who have requested your participation will not receive any personal information about you.

---

5781   **Consent and debriefing.** Because online studies are typically fully au-  
5782   tomated, participants do not have a chance to interact with researchers  
5783   around consent and debriefing. Further, engagement with long con-  
5784   sent forms may be minimal. In our work, we have typically relied on  
5785   short consent statements such as the one from our class that is shown  
5786   in table 12.2. Similarly, debriefing often occurs through a set of pages  
5787   that summarize all components of the debriefing process (participation  
5788   gratitude, discussion of goals, explanation of deception if relevant, and  
5789   questions and clarification). Because these interactions are so short, it  
5790   is especially important to include contact information prominently so  
5791   that participants can follow up.

### 5792   12.2.3 When to collect data online?

5793   Online data collection is increasingly ubiquitous in the behavioral  
5794   sciences. Further, the web browser—alongside survey software like  
5795   Qualtrics or packages like jsPsych (Leeuw 2023)—can be a major aid to  
5796   transparency in sharing experimental materials. Replication and reuse  
5797   of experimental materials is vastly simpler if readers and reviewers  
5798   can click a link and share the same experience as a participant in your  
5799   experiment. By and large, well-designed studies yield data that are  
5800   as reliable as in-lab data (M. Buhrmester, Kwang, and Gosling 2011;  
5801   Mason and Suri 2012; Crump, McDonnell, and Gureckis 2013).

5802 Still, online data collection is not right for every experiment. Studies  
5803 that have substantial deception or induce negative emotions may require  
5804 an experimenter present to alleviate ethical concerns or provide debrief-  
5805 ing. Beyond ethical issues, we discuss four broader concerns to consider  
5806 when deciding whether to conduct data collection online: (1) popula-  
5807 tion availability, (2) the availability of particular measures, (3) the feasi-  
5808 bility of particular manipulations, and (4) the length of experiments.

5809 **Population.** Not every target population can be tested online. Indeed,  
5810 initially, convenience samples from Amazon Mechanical Turk were the  
5811 only group easily available for online studies. More recently, new tools  
5812 have emerged to allow prescreening of crowd participants, including  
5813 sites like Cloud Research and Prolific (Eyal et al. 2021; Peer et al.  
5814 2021).<sup>10</sup> And it may initially have seemed implausible that children  
5815 could be recruited online, but during the COVID-19 pandemic a  
5816 substantial amount of developmental data collection moved online,  
5817 with many studies yielding comparable results to in-lab studies (e.g.,  
5818 Chuey et al. 2021).<sup>11</sup> Finally, new, non-US crowdsourcing platforms  
5819 continue to grow in popularity, leading to greater global diversity in  
5820 the available online populations.

5821 **Online measures.** Not all measures are available online, though more  
5822 and more are. Although online data collection was initially restricted

<sup>10</sup> These tools still have significant weaknesses for accessing socio-demographically diverse populations within and outside the US, however—screening tools can remove participants, but if the underlying population does not contain many participants from a particular demographic, it can be hard to gather large enough samples. For an example of using crowdsourcing and social media sites to gather diverse participants, see DeMayo et al. (2021).

<sup>11</sup> Sites like LookIt (<https://lookit.mit.edu>) now offer sophisticated platforms for hosting studies for children and families (Scott and Schulz 2017).

5823 to the use of survey measures—including ratings and text responses—  
5824 measurement options have rapidly expanded. The widespread use  
5825 of libraries like jsPsych (De Leeuw 2015) has meant that millisecond  
5826 accuracy in capturing response times is now possible within web  
5827 browsers; thus, most reaction time tasks are quite feasible (Crump,  
5828 McDonnell, and Gureckis 2013). The capture of sound and video is  
5829 possible with modern browser frameworks (Scott and Schulz 2017).  
5830 Further, even measures like mouse- and eye-tracking are beginning  
5831 to become available (Maldonado, Dunbar, and Chemla 2019; Slim  
5832 and Hartsuiker 2023). In general, almost any variable that can be  
5833 measured in the lab without specialized apparatus can also be collected  
5834 online. On the other hand, studies that measure a broader range of  
5835 physiological variables (e.g., heart rate or skin conductance) or a larger  
5836 range of physical behaviors (e.g., walking speed or pose) are still likely  
5837 difficult to implement online.

5838 **Online manipulations.** Online experiments are limited to the set of ma-  
5839 nipulations that can be created within a browser window—but this re-  
5840 striction excludes many different manipulations that involve real-time  
5841 social interactions with a human being.<sup>12</sup> Synchronous chat sessions can  
5842 be a useful substitute (Hawkins, Frank, and Goodman 2020), but these  
5843 focus the experiment on the content of what is said and exclude the  
5844 broader set of nonverbal cues available to participants in a live interac-

<sup>12</sup> So-called moderated experiments—in which the experimental session is administered through a synchronous video chat—have been used widely in online experiments for children, but these designs are less common in experiments with adults because they are expensive and time-consuming to administer (Chuey et al. 2021).

5845 tion (e.g., gaze, race, appearance, accent, and so on). Creative experi-  
5846 menters can circumvent these limitations by using pictures, videos, and  
5847 other methods. But more broadly, an experimenter interested in imple-  
5848 menting a particular manipulation online should ask how compelling  
5849 the online implementation is compared with an in-lab implementation.  
5850 If the intention is to induce some psychological state—say stress, fear, or  
5851 disgust—experimenters must trade off the greater ease of recruitment  
5852 and larger scale of online studies with the more compelling experience  
5853 they may be able to offer in a controlled lab context.

5854 **The length of online studies.** One last concern is about attention and  
5855 focus in online studies. Early guidance around online studies tended to  
5856 focus on making studies short and easy, with the rationale that crowd-  
5857 sourcing workers were used to short jobs. Our sense is that this guidance  
5858 no longer holds. Increasingly, researchers are deploying long and com-  
5859 plex batteries of tasks to relatively good effect (e.g., [Enkavi et al. 2019](#))  
5860 and conducting repeated longitudinal sampling protocols (discussed in  
5861 depth in [Litman and Robinson 2020](#)). Rather than relying on hard-and-  
5862 fast rules about study length, a better approach for online testing is to  
5863 ensure that participants' experience is as smooth and compelling as pos-  
5864 sible. Under these conditions, if an experiment is viable in the lab, it is  
5865 likely viable online.

5866 Online testing tools continue to grow and change, but they are already  
5867 mature enough that using them should be part of most behavioral re-  
5868 searchers' basic toolkit.<sup>13</sup>

<sup>13</sup> It is of course important to keep in mind that if a person works part- or full-time on a crowdsourcing platform, they are not a representative sample of the broader national population. Unfortunately, similar caveats hold true for in-person convenience samples (see chapter 10). Ultimately, researchers must reason about what their generalization goal is and whether that goal is consistent with the samples they can access (online or otherwise).

### 5869 *12.3 Ensuring high-quality data*

5870 In the final section of this chapter, we review some key data collection  
5871 practices that can help researchers collect high-quality data while re-  
5872 specting our ethical obligations to participants. By “high quality,” here  
5873 we especially mean datasets that are uncontaminated by responses gen-  
5874 erated by misunderstanding of instructions, fatigue, incomprehension,  
5875 or intentional neglect of the experimental task.

5876 We'll begin by discussing the issue of pilot testing; we recommend a  
5877 systematic procedure for piloting that can maximize the chance of col-  
5878 lecting high-quality data. Next, we'll discuss the practice of checking  
5879 participants' comprehension and attention and what such checks should  
5880 and shouldn't be used for. Finally, we'll discuss the importance of main-  
5881 taining consistent data collection records.

5882 12.3.1 *Conduct effective pilot studies*

5883 A **pilot study** is a small study conducted before you collect your main  
5884 sample. The goal is to ensure smooth and successful data collection by  
5885 first checking if your experimental procedures and data collection work-  
5886 flow are working correctly. Pilot studies are also an opportunity to get  
5887 feedback from participants about their experience of the experimental  
5888 task; for example, is it too easy, too difficult, or too boring.

5889 Because pilot studies usually involve a small number of participants, they  
5890 are not a reliable indicator of the study results, such as the expected ef-  
5891 fect size or statistical significance (as we discussed in chapter 10). *Don't*  
5892 use pilots to check if your effect is present or to estimate an effect size  
5893 for power analysis. What pilots *can* do is tell you about whether your  
5894 experimental procedure is viable. For example, pilot studies can re-  
5895 veal:

- 5896 – if your code crashes under certain circumstances
- 5897 – if your instructions confuse a substantial portion of participants
- 5898 – if you have a very high dropout rate
- 5899 – if your data collection procedure fails to log variables of interest
- 5900 – if participants are disgruntled by the end of the experiment

5901 We recommend that all experimenters perform—at the very  
5902 minimum—two pilot studies before they launch a new experiment.<sup>14</sup>

5903 The first pilot, which we call your **non-naive participant pilot**, can make  
5904 use of participants who know the goals of the experiment and under-  
5905 stand the experimental manipulation—this could be a friend, collabo-  
5906 rator, colleague, or family member.<sup>15</sup> The goal of this pilot study is to  
5907 ensure that your experiment is comprehensible, that participants can  
5908 complete it, and that the data are logged appropriately. You must *ana-*  
5909 *lyze* the data from the non-naive pilot, at least to the point of checking  
5910 that the relevant data about each trial is logged.

5911 The second pilot, your **naive participant pilot**, should consist of a test of  
5912 a small set of participants recruited via the channel you plan to use for  
5913 your main study. The number of participants you should pilot depends  
5914 on the cost of the experiment in time, money, and opportunity as well  
5915 as its novelty. A brand new paradigm is likely more prone to error than a  
5916 tried and tested paradigm. For a short online survey-style experiment, a  
5917 pilot of 10–20 people is reasonable. A more time-consuming laboratory  
5918 study might require piloting just two or three people.<sup>16</sup>

5919 The goal of the naive pilot study is to understand properties of the par-  
5920 ticipant experience. Were participants confused? Did they withdraw  
5921 before the study finished? Even a small number of pilots can tell you

<sup>14</sup> We mean especially when deploying a new experimental paradigm or when collecting data from a new population. Once you have run many studies with a similar procedure and similar sample, extensive piloting is less important. Any time you change something, it's always good to run one or two pilots, though, just to check that you didn't inadvertently mess up your experiment.

<sup>15</sup> In a pinch you can even run yourself through the experiment a bunch of times (though this isn't preferable because you're likely to miss a lot of aspects of the experience that you are habituated to, especially if you've been debugging the experiment already).

5922 that your dropout rate is likely too high: for example, if five of ten pilot  
5923 participants withdraw, you likely need to reconsider aspects of your de-  
5924 sign. It's critical for your naive participant pilot that you debrief more  
5925 extensively with your participants. This debriefing often takes the form  
5926 of an interview questionnaire after the study is over. "What did you  
5927 think the study was about?" and "Is there any way we could improve  
5928 the experience of being in the study?" can be helpful questions. Often  
5929 this debriefing is more effective if it is interactive, so even if you are  
5930 running an online study you may want to find some way to chat with  
5931 your participants.

5932 Piloting—especially piloting with naive participants to optimize the par-  
5933 ticipant experience—is typically an iterative process. We frequently  
5934 launch an experiment for a naive pilot, then recognize from the data  
5935 or from participant feedback that the experience can be improved. We  
5936 make tweaks and pilot again. Be careful not to over-fit to small dif-  
5937 ferences in pilot data, however. Piloting should be more like work-  
5938 shopping a manuscript to remove typos than doing statistical analysis. If  
5939 someone has trouble understanding a particular sentence—whether in  
5940 your manuscript or in your experiment instructions—you should edit  
5941 to make it clearer!

<sup>16</sup> In the case of especially expensive experiments, it can be a dilemma whether to run a larger pilot to identify difficulties since such a pilot will be costly. In these cases, one possibility is to plan to include the pilot participants in the main dataset if no major procedural changes are required. In this case, it is helpful to preregister a contingent testing strategy to avoid introducing data-dependent bias (see chapter 11). For example, in a planned sample of 100 participants, you could preregister running 20 as a pilot sample with the stipulation that you will look only at their dropout rate—and not at any condition differences. Then the preregistration can state that, if the dropout rate is lower than 25%, you will collect the next 80 participants and analyze the whole dataset, including the initial pilot, but if dropout rate is higher than 25%, you will discard the pilot sample and make changes. This kind of strategy can help you split the difference between cautious piloting and conservation of rare or costly data.

 ACCIDENT REPORT*Data logging much?*

When Mike was in graduate school, his lab got a contract to test a very large group of participants in a battery of experiments, bringing them into the lab over the course of a series of intense bursts of participant testing. He got the opportunity to add an experiment to the battery, allowing him to test a much larger sample than resources would otherwise allow. He quickly coded up a new experiment as part of a series of ongoing studies and began deploying it, coming to the lab every weekend for several months to help move participants through the testing protocol. Eagerly opening up the data file to reap the reward of this hard work, he found that the condition variable was missing from the data files. Although the experimental manipulation had been deployed properly, there was no record of which condition each participant had been run in, and so the data were essentially worthless. Had he run a quick pilot (even with non-naive participants) and attempted to analyze the data, this error would have been detected and many hours of participant and experimenter effort would not have been lost.

5942

5943 12.3.1 *Measure participant compliance*

5944 You've constructed your experiment and piloted it. You are almost  
5945 ready to go—but there is one more family of tricks for helping to  
5946 achieve high-quality data: integrating measures of participant com-

pliance into your paradigm. Collecting data on compliance (whether participants followed the experimental procedures as expected) can help you quantify whether participants understood your task, engaged with your manipulation, and paid attention to the full experimental experience. These measures in turn can be used both to modify your experimental paradigm and to exclude specific participants that were especially noncompliant (Hauser, Ellsworth, and Gonzalez 2018; Ejelöv and Luke 2020).

Below we discuss four types of compliance checks: (1) passive measures, (2) comprehension checks, (3) manipulation checks, and (4) attention checks. Passive measures and comprehension checks are very helpful for enhancing data quality. Manipulation checks also often have a role to play. In contrast, we typically caution in the use of attention checks.

1. **Passive measures of compliance.** Even if you do not ask participants anything extra in an experiment, it is often possible to tell if they have engaged with the experimental procedure simply by how long it takes them to complete the experiment. If you see participants with completion times substantially above or below the median, there is a good chance that they are either multi-tasking or rushing through the experiment without engaging.<sup>17</sup> Passive measures cost little to implement and should be inserted

<sup>17</sup> Measurements of per-page or per-element completion times can be even more specific since they can, for example, identify participants that simply did not read an assigned passage.

5968 whenever possible in experiments.<sup>18</sup>

5969       **2. Comprehension checks.** For tasks with complex instructions or  
5970        experimental materials (say a passage that must be understood for  
5971        a judgment to be made about it), it can be very helpful to get  
5972        a signal that participants have understood what they have read  
5973        or viewed. Comprehension checks, which ask about the content  
5974        of the experimental instructions or materials, are often included  
5975        for this purpose. For the comprehension of instructions, the best  
5976        kinds of questions simply query the knowledge necessary to suc-  
5977        ceed in the experiment: for example, “What are you supposed to  
5978        do when you see a red circle flash on the screen?” In many plat-  
5979        forms, it is possible to make participants reread the instructions  
5980        again until they can answer these correctly. This kind of repeti-  
5981        tion is nice because it corrects participants’ misconceptions rather  
5982        than allowing them to continue in the experiment when they do  
5983        not understand.<sup>19</sup>

5984       **3. Manipulation checks.** If your experiment involves more than  
5985        a very transient manipulation—for example, if you plan to in-  
5986        duce some state in participants or have them learn some content—  
5987        then you can include a measure in your experiment that confirms  
5988        that your manipulation succeeded (Ejelöv and Luke 2020). This

<sup>18</sup> One variation that we endorse in certain cases is to force participants to engage with particular pages for a certain amount of time through the use of timers. Though, beware, this kind of feature can lead to an adversarial relationship with participants—in the face of this kind of coercion, many will opt to pull out their phone and multi-task until the timer runs down.

<sup>19</sup> If you are querying comprehension of experimental materials rather than instructions, you may not want to reexpose participants to the same passage again in order to avoid confounding a participants’ initial comprehension and the amount of exposure that they receive.

measure is known as a manipulation check because it measures some prerequisite difference between conditions that is not the key causal effect of interest but is causally prerequisite to this effect. For example, if you want to see if anger affects moral judgment, then it makes sense to measure whether participants in your anger induction condition rate themselves as angrier than participants in your control condition. Manipulation checks are useful in the interpretation of experimental findings because they can decouple the failure of a manipulation from the failure of a manipulation to affect your specific measure of interest.<sup>20</sup>

4. **Attention checks.** A final type of compliance check is a check that participants are paying attention to the experiment at all. One simple technique is to add questions that have a known and fairly obvious right answer (e.g., “What’s the capital of the United States?”). These trials can catch participants that are simply ignoring all text and “mashing buttons,” but they will not find participants who are mildly inattentive. Sometimes experimenters also use trickier compliance checks, such as putting an instruction for participants to click a particular answer deep within a question text that otherwise would have a different answer (figure 12.2). Such compliance checks decrease so-called satisficing behavior, in which participants read as quickly as they

<sup>20</sup> Hauser, Ellsworth, and Gonzalez (2018) worry that manipulation checks can themselves change the effect of a manipulation—this worry strikes us as sensible, especially for some types of manipulations like emotion inductions. Their recommendation is to test the efficacy of the manipulation in a separate study, rather than trying to nest the manipulation check within the main study.

6011 can to get away with doing only the minimum. On the other  
 6012 hand, participants may see such trials as indications that the  
 6013 experimenter is trying to trick them, and adopt a more adversarial  
 6014 stance toward the experiment, which may result in less compli-  
 6015 ance with other aspects of the design, unless they are at the end  
 6016 of the experiment (Hauser, Ellsworth, and Gonzalez 2018). If  
 6017 you choose to include attention checks like these, be aware that  
 6018 you are likely reducing variability in your sample—trading off  
 6019 representativeness for compliance.

#### Sports Participation

Most modern theories of decision making recognize the fact that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. In order to facilitate our research on decision making we are interested in knowing certain factors about you, the decision maker. Specifically, we are interested in whether you actually take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions will be ineffective. So, in order to demonstrate that you have read the instructions, please ignore the sports items below, as well as the continue button. Instead, simply click on the title at the top of this screen (i.e., "sports participation") to proceed to the next screen.  
 Thank you very much.

Which of these activities do you engage in regularly?  
 (click on all that apply)

skiing	soccer	snowboarding	running	hockey
football	swimming	tennis	basketball	cycling

**Continue**

Figure 12.2

An attention check trial based on Oppenheimer, Meyvis, and Davidenko (2009). These trials can decrease variability in participant attention, but at the cost of selecting a subsample of participants, so they should be used cautiously.

6020 Data from all of these types of checks are used in many different—often  
 6021 inconsistent—ways in the literature. We recommend that you:

- 6022 1. Use passive measures and comprehension checks as preregistered  
 6023 exclusion criteria to eliminate a (hopefully small) group of partic-  
 6024 ipants who might be noncompliant with your experiment.

6025        2. Check that exclusions are low and that they are uniform across  
6026        conditions. If exclusion rates are high, your design may have  
6027        deeper issues. If exclusions are asymmetric across conditions, you  
6028        may be compromising your randomization by creating a situation  
6029        in which (on average) different kinds of participants are included  
6030        in one condition compared with the other. Both of these situa-  
6031        tions substantially compromise any estimate of the causal effect of  
6032        interest.

### ⚠️ ACCIDENT REPORT

#### *Does data quality vary throughout the semester?*

Every lab that collects empirical data repeatedly using the same popula-  
tion builds up lore about how that population varies in different contexts.

Many researchers who conducted experiments with college undergrad-  
uates were taught never to run their studies at the end of the semester.  
Exhausted and stressed students would likely yield low-quality data, or so  
the argument went. Until the rise of multi-lab collaborative projects like  
ManyLabs (see chapter 3), such beliefs were almost impossible to test.

ManyLabs 3 aimed specifically to evaluate data quality variation across  
the academic calendar (Ebersole et al. 2016). With 2,696 participants  
at 20 sites, the study conducted replications of 13 previously published  
findings. Although only six of these findings showed strong evidence of  
replicating across sites, none of the six effects was substantially moderated

by being collected later in the semester. The biggest effect they observed was a change in the Stroop effect from  $d = 0.89$  during the beginning and middle of the semester to  $d = 0.92$  at the end. There was some evidence that participants *reported* being less attentive at the end of the semester, but this trend wasn't accompanied by a moderation of experimental effects.

Researchers are subject to the same cognitive illusions and biases as any human. One of these biases is the search to find meaning in the random fluctuations they sometimes observe in their experiments. The intuitions formed through this process can be helpful prompts for generating hypotheses—but beware of adopting them into your “standard operating procedures” without further examination. Labs that avoided data collection during the end of the semester might have sacrificed 10%–20% of their data collection capacity for no reason!

6034

6035 3. Deploy manipulation checks if you are concerned about whether  
6036 your manipulation effectively induces a difference between  
6037 groups. Analyze the manipulation check separately from the  
6038 dependent variable to test whether the manipulation was causally  
6039 effective (Ejelöv and Luke 2020).

6040 4. Make sure that your attention checks are not confounded in any  
6041 way with condition—remember our cautionary tale from chap-  
6042 ter 9, in which an attention check that was different across condi-  
6043 tions actually created an experimental effect.

6044        5. *Do not* include any of these checks in your analytic models as a  
6045           covariate, as including this information in your analysis compro-  
6046           mises the causal inference from randomization and introduces bias  
6047           in your analysis (Montgomery, Nyhan, and Torres 2018).<sup>21</sup>

6048        Used appropriately, compliance checks can provide both a useful set  
6049           of exclusion criteria and a powerful tool for diagnosing potential issues  
6050           with your experiment during data analysis and correcting them down  
6051           the road.

<sup>21</sup> Including this information means you are “conditioning on a post-treatment variable,” as we described in chapter 7. In medicine, analysts distinguish “intent-to-treat” analysis, where you analyze data from everyone you gave a drug, and “as-treated” analysis, where you analyze data depending on how much of the drug people actually took. In general, intent-to-treat analysis gives you the generalizable causal estimate. In our current situation, if you include compliance as a covariate, you are essentially doing an “as-treated” analysis and your estimate can be biased as a result. Although there is occasional need for such analyses, in general you probably want to avoid them.

6052 *12.3.1 Keep consistent data collection records*

6053 As an experimentalist, one of the worst feelings is to come back to  
 6054 your data directory and see a group of data files, `run1.csv`, `run2.csv`,  
 6055 `run3.csv` and not know what experimental protocol was run for each.  
 6056 Was `run1` the pilot? Maybe a little bit of personal archaeology with  
 6057 timestamps and version history can tell you the answer, but there is no  
 6058 guarantee.<sup>22</sup>

<sup>22</sup> We'll have a lot to say about this issue in chapter 13.

	A	B	C	D	E	F	G
1	DOT	RA	SID	DOB	Gender	Consent	Source
2	12/14/12	ak, fp	ASD_01	9/19/98	m	1	fp
3	12/17/12	ak, fp	ASD_02	6/17/90	f	0	fp
4	12/18/12	ak, fp	ASD_03	8/15/90	f	1	fp
5	12/20/12	mf, fp	ASD_04	9/21/08	m	1	fp
6	1/21/13	mf, fp	ASD_05	8/31/07	m	1	fp
7	1/29/13	ak, ca	ASD_06	8/30/10	f	1	ah
8	1/31/13	ak, fp	ASD_07	10/26/05	m	1	fp

6059 As well as collecting the actual data in whatever form they take (e.g.,  
 6060 paper surveys, videos, or files on a computer), it is important to log  
 6061 metadata—data about your data—including relevant information like  
 6062 the date of data collection, the sample that was collected, the experi-  
 6063 ment version, and the research assistants who were present. The rele-  
 6064 vant metadata will vary substantially from study to study—the impor-  
 6065 tant part is that you keep detailed records. Figure 12.3 and figure 12.4  
 6066 give two examples from our own research. The key feature is that they  
 6067 provide some persistent metadata about how the experiments were con-  
 6068 ducted.

Figure 12.3  
 Part of a run sheet for a developmental study.

---

```
%%%%%%
Added a simple familiarization slide substitute that presents Bob and
shows that the experiment is about a person talking to you. Before
that, the familiarization slide was simply skipped.
%%%%%
```

---

```
-----  
November 18 2013  
50 subjects | Betting | No familiarization | Friend  
var participant_response_type = 1;  
var participant_feature_count = 1;  
var linguistic_framing = 0;  
var question_type = 0;  
-----  
November 18 2013  
50 subjects | Likert | No familiarization | Friend  
var participant_response_type = 2;  
var participant_feature_count = 1;  
var linguistic_framing = 0;  
var question_type = 2;  
%%%%%
The experiment now asked the subjects the referent of Bobs statement
at the bottom of the page. The previous experiments always had the
input field just below the stimuli or, in the case of 3fc hoovering
over the images did highlighted possible ones.
%%%%%
```

---

```
-----  
November 30 2013 ~ 7 pm:  
50 subjects | 3 forced choice condition | No familiarization | Friend  
var participant_response_type = 0;  
var participant_feature_count = 1;  
var linguistic_framing = 0;  
var question_type = 0;
```

---

Figure 12.4  
An excerpt of a log for an iterative run of online experiments.

## 6069 12.4 Chapter summary: Data collection

6070 In this chapter, we took the perspective of both the participant and the  
6071 researcher. Our goal was to discuss how to achieve a good research  
6072 outcome for both. On the side of the participant, we highlighted the  
6073 responsibility of the experimenter to ensure a robust consent and de-  
6074 briefing process. We also discussed the importance of a good experi-  
6075 mental experience in the lab and online—ensuring that the experiment  
6076 is not only conducted ethically but is also pleasant to participate in. Fi-  
6077 nally, we discussed how to address some concerns about data quality  
6078 from the researcher perspective, recommending both the extensive use  
6079 of non-naive and naive pilot participants and the use of comprehension  
6080 and manipulation checks.



### DISCUSSION QUESTIONS

1. “Citizen science” is a movement to have a broader base of individuals participate in research because they are interested in discoveries and want to help. In practice, citizen science projects in psychology like Project Implicit (<https://implicit.harvard.edu/implicit>), Children Helping Science (<https://lookit.mit.edu>), and TheMusicLab.org (<https://themusiclab.org>) have all succeeded by offering participants a compelling experience. Check one of these out, participate in a study, and make a list of the features that make it fun and easy to contribute data.

2. Be a Turker! Sign up for an account as an Amazon Mechanical Turk or Prolific Academic worker and complete a couple of tasks. How did you feel about browsing the list of tasks looking for work? What features of tasks attracted your interest? How hard was it to figure out how to participate in each task? And how long did it take to get paid?

6082

## READINGS

- An introduction to online research: Buhrmester, Michael D., Sanaz Talaifar, and Samuel D. Gosling (2018). “An Evaluation of Amazon’s Mechanical Turk, Its Rapid Rise, and Its Effective Use.” *Perspectives on Psychological Science* 13 (2): 149–154. <https://doi.org/10.1177/1745691617706516>.

6083

## References

Allen, Michael. 2017. “Debriefing of Participants.” In *The SAGE Encyclopedia of Communication Research Methods*. Vol. 1–4. Thousand Oaks, CA: Sage Publications.

Anderson, Craig A., Johnie J. Allen, Courtney Plante, Adele Quigley-McBride, Alison Lovett, and Jeffrey N. Rokkum. 2019. “The MTurkification of Social and Personality Psychology.” *Personality and Social Psychology Bulletin* 45 (6): 842–850.

Benjamin, Ludy T. 2000. “The Psychology Laboratory at the Turn of the 20th Century.” *American Psychologist* 55 (3): 318–321. <https://doi.org/10.1037/0003-066X.55.3.318>.

6085

- Buhrmester, Michael D., Sanaz Talaifar, and Samuel D. Gosling. 2018. “An Evaluation of Amazon’s Mechanical Turk, Its Rapid Rise, and Its Effective Use.” *Perspectives on Psychological Science* 13 (2): 149–154. <https://doi.org/10.1177/1745691617706516>.
- Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. “Amazon’s Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?” *Perspectives on Psychological Science* 6 (1): 3–5.
- Chuey, Aaron, Mika Asaba, Sophie Bridgers, Brandon Carrillo, Griffin Dietz, Teresa Garcia, Julia A. Leonard, et al. 2021. “Moderated Online Data-Collection for Developmental Research: Methods and Replications.” *Frontiers in Psychology*, 4968.
- Cialdini, Robert B., and Noah J. Goldstein. 2004. “Social Influence: Compliance and Conformity.” *Annual Review of Psychology* 55 (1): 591–621.
- Crump, Matthew J. C., John V. McDonnell, and Todd M. Gureckis. 2013. “Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research.” *PLoS One* 8 (3): e57410.
- De Leeuw, Joshua R. 2015. “jsPsych: A JavaScript Library for Creating Behavioral Experiments in a Web Browser.” *Behavior Research Methods* 47 (1): 1–12.
- DeMayo, Benjamin, Danielle Kellier, Mika Braginsky, Christina Bergmann, Cielke Hendriks, Caroline F. Rowland, Michael Frank, and Virginia Marchman. 2021. “Web-CDI: A System for Online Administration of the MacArthur-Bates Communicative Development Inventories.” *Language Development Research* 1 (1): 55–98. <https://doi.org/10.34842/kr8e-w591>.
- Ebersole, Charles R., Olivia E. Atherton, Aimee L. Belanger, Hayley M. Skul-

- borstad, Jill M. Allen, Jonathan B. Banks, Erica Baranski, et al. 2016. “Many Labs 3: Evaluating Participant Pool Quality Across the Academic Semester via Replication.” *Journal of Experimental Social Psychology* 67 (November 2016): 68–82. <https://doi.org/10.1016/j.jesp.2015.10.012>.
- Ejelöv, Emma, and Timothy J. Luke. 2020. “‘Rarely Safe to Assume’: Evaluating the Use and Interpretation of Manipulation Checks in Experimental Social Psychology.” *Journal of Experimental Social Psychology* 87 (2020): 103937.
- Enkavi, A. Zeynep, Ian W. Eisenberg, Patrick G. Bissett, Gina L. Mazza, David P. MacKinnon, Lisa A. Marsch, and Russell A. Poldrack. 2019. “Large-Scale Analysis of Test–Retest Reliabilities of Self-Regulation Measures.” *Proceedings of the National Academy of Sciences* 116 (12): 5472–5477.
- Eyal, Peer, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. “Data Quality of Platforms and Panels for Online Behavioral Research.” *Behavior Research Methods* 54 (2022): 1643–1662.
- Fisher, Jill A. 2013. “Expanding the Frame of ‘Voluntariness’ in Informed Consent: Structural Coercion and the Power of Social and Economic Context.” *Kennedy Institute of Ethics Journal* 23 (4): 355–379.
- Fitzpatrick, Emily F. M., Alexandra L. C. Martiniuk, Heather D’Antoine, June Oscar, Maureen Carter, and Elizabeth J. Elliott. 2016. “Seeking Consent for Research with Indigenous Communities: A Systematic Review.” *BMC Medical Ethics* 17 (1): 1–18.
- Gass, Robert H., and John S. Seiter. 2018. *Persuasion: Social Influence and Compliance Gaining*. Routledge.
- Gramlich, John. 2021. “America’s Incarceration Rate Falls to Lowest Level

- Since 1995.” Pew Research Center. <https://pewrsr.ch/2rfSmVL>.
- Hara, Kotaro, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. “A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3173574.3174023>.
- Hauser, David J., Phoebe C. Ellsworth, and Richard Gonzalez. 2018. “Are Manipulation Checks Necessary?” *Frontiers in Psychology* 9:998. <https://doi.org/10.3389/fpsyg.2018.00998>.
- Hawkins, Robert D., Michael C. Frank, and Noah D. Goodman. 2020. “Characterizing the Dynamics of Learning in Repeated Reference Games.” *Cognitive Science* 44 (6): e12845.
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?” *Behavioral and Brain Sciences* 33 (2-3): 61–83.
- Holmes, David S. 1976. “Debriefing after Psychological Experiments: I. Effectiveness of Postdeception Dehoaxing.” *American Psychologist* 31 (12): 858–867. <https://doi.org/10.1037/0003-066X.31.12.858>.
- Irani, Lilly C., and M. Six Silberman. 2013. “Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 611–620.
- Kadam, Rashmi Ashish. 2017. “Informed Consent Process: A Step Further towards Making It Meaningful!” *Perspectives in Clinical Research* 8 (3): 107–112.
- Leeuw, Gilbert de, Joshua R. 2023. “jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments.” *Journal of Open Source Software* 6088.

- Software 8 (85): 5351. <https://doi.org/10.21105/joss.05351>.
- Litman, Leib, and Jonathan Robinson. 2020. *Conducting Online Research on Amazon Mechanical Turk and Beyond*. Sage Publications.
- Litman, Leib, Jonathan Robinson, and Tzvi Abberbock. 2017. “TurkPrime. Com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences.” *Behavior Research Methods* 49 (2): 433–442.
- Maldonado, Mora, Ewan Dunbar, and Emmanuel Chemla. 2019. “Mouse Tracking as a Window into Decision Making.” *Behavior Research Methods* 51 (3): 1085–1101.
- Mason, Winter, and Siddharth Suri. 2012. “Conducting Behavioral Research on Amazon’s Mechanical Turk.” *Behavior Research Methods* 44 (1): 1–23.
- Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. “How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It.” *American Journal of Political Science* 62 (3): 760–775.
- Moss, Aaron J., Cheskie Rosenzweig, Jonathan Robinson, and Leib Litman. 2020. “Demographic Stability on Mechanical Turk Despite COVID-19.” *Trends in Cognitive Sciences* 24 (9): 678–680.
- Office for Human Research Protections. 2003. “Prisoner Involvement in Research.” <https://www.hhs.gov/ohrp/regulations-and-policy/guidance/prisoner-research-ohrp-guidance-2003/index.html>.
- Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko. 2009. “Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power.” *Journal of Experimental Social Psychology* 45 (4): 867–872.
- Peer, Eyal, David M. Rothschild, Zak Evernden, Andrew Gordon, and Ekaterina Damer. 2021. “MTurk, Prolific or Panels? Choosing the Right

- Audience for Online Research.” *Social Science Research Network*. <https://doi.org/10.2139/SSRN.3765448>.
- Salehi, Niloufar, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. 2015. “We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers.” In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1621–1630.
- Scott, Kimberly, and Laura Schulz. 2017. “Lookit (Part 1): A New Online Platform for Developmental Research.” *Open Mind* 1 (1): 4–14.
- Sears, David O. 1986. “College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology’s View of Human Nature.” *Journal of Personality and Social Psychology* 51 (3): 515–530.
- Sieber, Joan E., and Michael J. Saks. 1989. “A Census of Subject Pool Characteristics and Policies.” *American Psychologist* 44 (7): 1053–1061.
- Slim, Mieke Sarah, and Robert J. Hartsuiker. 2023. “Moving Visual World Experiments Online? A Web-Based Replication of Dijkgraaf, Hartsuiker, and Duyck (2017) Using PCIbex and WebGazer. Js.” *Behavior Research Methods* 55 (7): 3786–3804.
- Young, Daniel R., Donald T. Hooker, and Fred E. Freeberg. 1990. “Informed Consent Documents: Increasing Comprehension by Reducing Reading Level.” *IRB: Ethics & Human Research* 12 (3): 1–5.

## 6091 13 PROJECT MANAGEMENT

### LEARNING GOALS

- Manage your research projects efficiently and transparently
- Develop strategies for data organization
- Optimize sharing of research products, like data and analysis code, by ensuring they are findable, accessible, interoperable, reusable (FAIR)
- Discuss potential ethical constraints on sharing research products

6092

6093 Your closest collaborator is you six months ago, but you  
6094 don't reply to emails.

6095 –Karl Broman (2015), quoting @gonuke<sup>1</sup>

<sup>1</sup> <https://twitter.com/gonuke>

6096 Have you ever returned to an old project folder to find a chaotic mess  
6097 of files with names like analysis-FINAL, analysis-FINAL-COPY, and  
6098 analysis-FINAL-COPY-v2? Which file is actually the final version!?  
6099 Or perhaps you've spent hours searching for a data file to send to your

6100 advisor, only to realize with horror that it was *only* stored on your old  
6101 laptop—the one that experienced a catastrophic hard drive failure when  
6102 you spilled coffee all over it one sleepy Sunday morning. These experi-  
6103 ences may make you sympathetic to Karl Broman’s quip above. Good  
6104 project management practices not only make it easier to share your re-  
6105 search with others, they also make for a more efficient and less error-  
6106 prone workflow that will avoid giving your future self a headache. This  
6107 chapter is about the process of managing all of the products of your re-  
6108 search workflow—methodological protocols, materials,<sup>2</sup> data, and anal-  
6109 ysis scripts. We focus especially on managing projects in ways that max-  
6110 imize their value to you and to the broader research community by  
6111 aligning with open science practices (maximizing TRANSPARENCY).

6112 When we talk about research products, we typically think of articles  
6113 in academic journals, which have been scientists’ main method of com-  
6114 munication since the scientific revolution in the 1600s.<sup>3</sup> But articles  
6115 only provide written summaries of research; they are not the original  
6116 research products. In recent years, there have been widespread calls  
6117 for increased sharing of research products, such as materials, data, and  
6118 analysis code (Munafò et al. 2017). When shared appropriately, these  
6119 other products can be as valuable as a summary article: Shared stimulus  
6120 materials can be reused for new studies in creative ways; shared anal-  
6121 ysis scripts can allow for reproduction of reported results and become

<sup>2</sup> We use the term “materials” here to cover a range of things another researcher might need in order to repeat your study—for example, stimuli, survey instruments, and code for computer-based experiments.

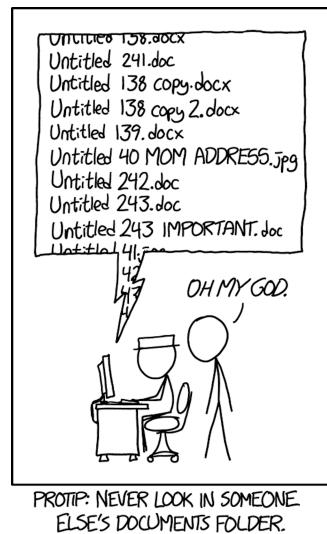


Figure 13.1

Poor file management creates chaos! “Documents” by xkcd (<https://xkcd.com/1459>, licensed under <https://xkcd.com/license.html>).

<sup>3</sup> The world’s oldest scientific journal is the *Philosophical Transactions of the Royal Society*, first published in 1665.

6122 templates for new analyses; and shared data can enable new analyses or  
6123 meta-analyses. Indeed, many funding agencies, and some journals, now  
6124 require that research products be shared publicly, except when there  
6125 are justified ethical or legal constraints, such as with sensitive medical  
6126 data (Nosek et al. 2015).

6127 Data sharing, in particular, has been the focus of intense interest. Shar-  
6128 ing data is associated with benefits in terms of error detection (Hard-  
6129 wicke et al. 2021), creative reuse that generates new discoveries (Voytek  
6130 2016), increased citations (Piwowar and Vision 2013), and detection of  
6131 fraud (Simonsohn 2013). According to surveys, researchers are usually  
6132 willing to share data in principle (Houtkoop et al. 2018), but unfor-  
6133 tunately, in practice, they often do not, even if you directly ask them  
6134 (Hardwicke and Ioannidis 2018). Often authors simply do not respond,  
6135 but when they do, they frequently report that data have been lost be-  
6136 cause they were stored on a misplaced or damaged computer or drive, or  
6137 team members with access to the data are no longer contactable (Tenopir  
6138 et al. 2020).

6139 As we have discussed in chapter 3, even when data are shared, they are  
6140 not always formatted in a way that they can be easily understood and  
6141 reused by other researchers, or even the original authors! This issue  
6142 highlights the critical role of **metadata**: information that documents

6143 the data (and other products) that you share, including README files,  
 6144 **codebooks** that document datasets themselves, and licenses that provide  
 6145 legal restrictions on reuse. We will discuss best practices for metadata  
 6146 throughout the chapter.

6147 Sound project management practices and sharing of research projects  
 6148 are mutually reinforcing goals that bring benefits for both yourself, the  
 6149 broader research community, and scientific progress. One particularly  
 6150 important benefit of good project management practices is that they en-  
 6151 able reproducibility. As we discussed in chapter 3, computational repro-  
 6152 ducibility involves being able to trace the provenance of any reported an-  
 6153 alytic result in a research report back to its original source. That means  
 6154 being able to recreate the entire analytic chain from data collection to  
 6155 data files, though analytic specifications to the research results reported  
 6156 in text, tables, and figures. If data collection is documented appropri-  
 6157 ately, and if data are stored, organized, and shared, then the provenance  
 6158 of a particular result is relatively easy to verify. But once this chain (fig-  
 6159 ure 13.2) is broken, it can be hard to reconstruct (Hardwicke et al. 2018).  
 6160 That's why it's critical to build good project management practices into  
 6161 your research workflow right from the start.

6162 In this chapter, you will learn how to manage your research project both  
 6163 efficiently and transparently.<sup>4</sup> Working toward these goals can create a

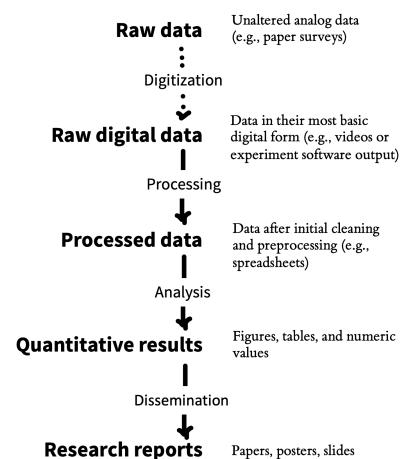


Figure 13.2  
 An illustration of the analytic chain from raw data through to research report.

<sup>4</sup> This chapter—especially the last section—draws heavily on Klein et al. (2018), an article on research transparency that several of us contributed to.

6164 virtuous cycle: if you organize your research products well, they are eas-  
6165 ier to share later, and if you assume that you will be sharing, you will be  
6166 motivated to organize your work better! We begin by discussing some  
6167 important principles of project management, including folder structure,  
6168 file naming, organization, and version control. Then we zoom in specif-  
6169 ically on data and discuss best practices for data sharing. We end by  
6170 discussing the question of what research products to share and some  
6171 of the potential ethical issues that might limit your ability to share in  
6172 certain circumstances.



### CASE STUDY

#### *ManyBabies, ManySpreadsheetFormats!*

The ManyBabies project is an example of “Big Team Science” in psychology. A group of developmental psychology researchers (including some of us) were worried about many of the issues of reproducibility, replicability, and experimental methods that we’ve been discussing throughout this book, so they set up a large-scale collaboration to replicate key effects in developmental science. The first of these studies was ManyBabies 1 (The ManyBabies Consortium et al. 2020), a study of infants’ preference for baby-talk (also known as “infant directed speech”).

The core team expected a handful of labs to contribute, but after a year-long data collection period, they ended up receiving data from 69 labs around the world! The outpouring of interest signaled a lot of enthusi-

asm from the community for this kind of collaborative science. Unfortunately, it also made for a tremendous data management headache. All kinds of complications and hilarity ensued as the idiosyncratic data formatting preferences of the various labs were reorganized to fit into a single standardized analysis pipeline (Byers-Heinlein et al. 2020).

All of the specific formatting changes that individual labs made were reasonable—altering column names for clarity, combining templates into a single Excel file, changing units (e.g., from seconds to milliseconds)—but together they created a very challenging **data validation** problem for the core analysis team, requiring many dozens of hours of coding and hand-checking. The data checking was critical: an error in one lab’s data was flagged during validation and led to the painful decision to drop those data from the final dataset. In future ManyBabies projects, the group has committed to using shared data validation software (<https://manybabies.org/validator>) to ensure that data files uploaded by individual labs conform to a shared standard.

6174

### 6175 *13.1 Principles of project management*

6176 A lot of project management problems can be avoided by following a  
6177 very simple file organization system.<sup>5</sup> For those researchers who “grew  
6178 up” managing their files locally on their own computers and email-  
6179 ing colleagues versions of data files and manuscripts with names like

6180 manuscript-FINAL-JS-rev1.xlsx, a few aspects of this system may  
6181 seem disconcerting. However, with a little practice, this new way of  
6182 working will start to feel intuitive and have substantial benefits.

6183 Here are the principles:

6184 1. There should be exactly one definitive copy of each document  
6185 in the project, with its name denoting what it is. For example,

6186 fifo\_manuscript.Rmd or fifo\_manuscript.docx is the write-  
6187 up of the “fifo” project as a journal manuscript.

6188 2. The location of each document should be within a folder that  
6189 serves to uniquely identify the document’s function within the  
6190 project. For example,

6191 analysis/experiment1/eye\_tracking\_preprocessing.Rmd  
6192 is clearly the file that performs preprocessing for the analysis of  
6193 eye-tracking data from experiment 1.

6194 3. The full project should be accessible to all collaborators via the  
6195 cloud, either using a version control platform (e.g., GitHub) or  
6196 another cloud storage provider (e.g., Dropbox, Google Drive).

6197 4. The revision history of all text and text-based documents  
6198 (minimally, data, analysis code, and manuscript files) should be  
6199 archived automatically. Automatic versioning is the key feature  
6200 of all version control systems and is often included by cloud

6201 storage providers.

6202 Keeping these principles in mind, we discuss best practices for project  
6203 organization, version control, and file naming.

6204 *13.1.1 Organizing your project*

6205 To the greatest extent possible, all files related to a project should be  
6206 stored in the same project folder (with appropriate subfolders), and on  
6207 the same storage provider. There are cases where this is impractical  
6208 due to the limitations of different software packages. For example, in  
6209 many cases, a team will manage its data and analysis code via GitHub but  
6210 decide to write collaboratively using Google Docs, Overleaf, or another  
6211 collaborative platform. (It can also be hard to ask all collaborators to use  
6212 a version control system they are unfamiliar with.) In that case, the final  
6213 paper should still be linked in some way to the project repository.<sup>6</sup>

6214 Figure 13.3 shows an example project stored on the Open Science  
6215 Framework. The top-level folder contains subfolders for analyses,  
6216 materials, raw and processed data (kept separately). It also contains the  
6217 paper manuscript and, critically, a README file in a text format that  
6218 describes the project. A README is a great way to document any

<sup>6</sup> The biggest issue that comes up in using a split workflow like this is the need to ensure reproducible written products, a process we cover in chapter 14.

6219 other metadata that the authors would like to be associated with the  
 6220 research products, for example a license, explained below.

Name	Modified
Example project (/rpydu/)	
- OSF Storage (United States)	
+ Analyses	
Heycke, Aust, & Stahl (2017) Subliminal influence on prefer... 2018-01-12 06:29 AM	
+ Material	
+ Processed data	
+ Raw data	
README.md	2018-06-12 07:26 AM
Study protocol (Stage-1 registered report).pdf	2018-01-12 06:33 AM

Figure 13.3

Sample top-level folder structure for a project. From Klein et al. (2018). Original visible on the Open Science Framework (<https://osf.io/xf6ug>).

6221 There are many reasonable ways to organize the subfolders of a research  
 6222 project, but the broad categories of materials, data, analysis, and writing  
 6223 are typically present.<sup>7</sup> In some projects—such as those involving multi-  
 6224 ple experiments or complex data types—you may have to adopt a more  
 6225 complex structure. In many of our projects, it's not uncommon to find  
 6226 paths like `/data/raw_data/exp1/demographics`. The key principle  
 6227 is to create a hierarchical structure in which subfolders uniquely iden-  
 6228 tify the part of the broader space of research products that are found  
 6229 inside them—that is, `/data/raw_data/exp1` contains all the raw data  
 6230 from experiment 1, and `/data/raw_data/exp1/demographics` con-  
 6231 tains all the raw *demographics* data from that particular experiment.

<sup>7</sup> We like the scheme followed by Project TIER (<https://www.projecttier.org>), which provides very clear guidance about file structure and naming conventions. TIER is primarily designed for a copy-and-paste workflow, which is slightly different from the “dynamic documents” workflow that we primarily advocate for (e.g., using R Markdown or Quarto as in appendix C).

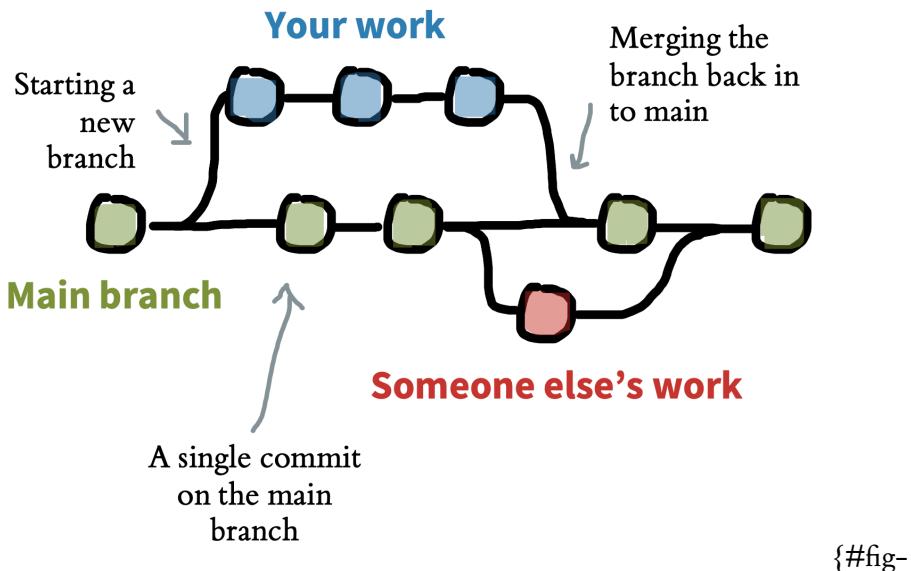
### 6232 13.1.2 Versioning

6233 Probably everyone who has ever collaborated electronically has experi-  
 6234 enced the frustration of editing a document, only to find out that you

6235 are editing the wrong version—perhaps some of the problems you are  
6236 working on have already been corrected, or perhaps the section you are  
6237 adding has already been written by someone else. A second common  
6238 source of frustration comes when you take a wrong turn in a project,  
6239 perhaps by reorganizing a manuscript in a way that doesn't work or  
6240 refactoring code in a way that turns out to be short-sighted.

6241 These two problems are solved by modern version control systems. Here  
6242 we focus on the use of **Git**, which is the most widely used version  
6243 control system. Git is a great general solution for version control, but  
6244 many people—including several of us—don't love it for collaborative  
6245 manuscript writing. We'll introduce Git and its principles here, while  
6246 noting that online collaboration tools like Google Docs and Overleaf<sup>8</sup>  
6247 can be easier for writing prose (as opposed to code); we cover this topic  
6248 in a bit more depth in chapter 14.

<sup>8</sup> Overleaf is actually supported by Git on the backend!



6249

{#fig-

6250 management-git .column-margin fig-alt="A diagram of connected  
 6251 circles where "your work" and "someone else's work" branch off of  
 6252 "main branch" then merge back in."}

6253 Git is a tool for creating and managing projects, which are called repos-  
 6254 itories. A Git repository is a directory whose revision history is tracked  
 6255 via a series of **commits**—snapshots of the state of the project. These  
 6256 commits can form a tree with different **branches**, as when two con-  
 6257 tributors to the project are working on two different parts simultane-  
 6258 ously (@fig-management-git). These branches can later be **merged** ei-  
 6259 ther automatically or via manual intervention in the case of conflicting  
 6260 changes.

6261 Commonly, Git repositories are hosted by an online service like  
 6262 GitHub<sup>9</sup> to facilitate collaboration. With this workflow, a user makes

<sup>9</sup> <https://github.com>

6263 changes to a local version of the repository on their own computer  
6264 and **pushes** those changes to the online repository. Another user can  
6265 then **pull** those changes from the online repository to their own local  
6266 version. The online “origin” copy is always the definitive copy of  
6267 the project, and a record is kept of all changes. Chapter B provides a  
6268 practical introduction to Git and GitHub, and there are a variety of  
6269 good tutorials available online and in print ([Blischak, Davenport, and](#)  
6270 [Wilson 2016](#)).

6271 Collaboration using version control tools is designed to solve many of  
6272 the problems we’ve been discussing:

- 6273 – A remotely hosted Git repository is a cloud-based backup of your  
6274 work, meaning it is less vulnerable to accidental erasure.<sup>10</sup>
- 6275 – By virtue of having versioning history, you have access to previous  
6276 drafts in case you find you have been following a blind alley and  
6277 want to roll back your changes.
- 6278 – By creating new branches, you can create another, parallel history  
6279 for your project so that you can try out major changes or additions  
6280 without disturbing the main branch in the process.
- 6281 – A project’s commit history is labeled with each commit’s author  
6282 and date, facilitating record-keeping and collaboration.

<sup>10</sup> In 48BC, Julius Caesar accidentally burned down part of the Great Library of Alexandria where the sole copies of many valuable ancient works were stored. To this day, many scientists have apparently retained the habit of storing single copies of important information in vulnerable locations. Even in the age of cloud computing, hard drive failure is a surprisingly common source of problems!

- 6283 – Automatic merging can allow synchronous editing of different  
6284 parts of a manuscript or codebase.<sup>11</sup>

6285 Organizing a project repository for collaboration and hosting on a re-  
6286 mote platform is an important first step toward sharing! Many of our  
6287 projects (like this book) are actually born open: we do all of our work  
6288 on a publicly hosted repository for everyone to see (Rouder 2015). This  
6289 philosophy of “working in the open” encourages good organization  
6290 practices from the beginning. It can feel uncomfortable at first, but this  
6291 discomfort soon vanishes as you realize that basically no one is looking  
6292 at your in-progress project.

6293 One concern that many people raise about sharing in-progress research  
6294 openly is the possibility of “scooping”—that is, other researchers getting  
6295 an idea or even data from the repository and writing a paper before  
6296 you do. We have two responses to this concern. First, the empirical  
6297 frequency of this sort of scooping is difficult to determine but likely very  
6298 low—we don’t know of any documented cases. Mostly, the problem is  
6299 getting people to care about your experiment at all, not people caring  
6300 so much that they would publish using your data or materials! In Gary  
6301 King’s words (King and Shieber 2013), “The thing that matters the least  
6302 is being scooped. The thing that matters the most is being ignored.”  
6303 On the other hand, if you are in an area of research that you perceive

<sup>11</sup> Version control isn’t magic, and if you and a collaborator edit the same paragraph or function, you will likely have to merge your changes by hand. But Git will at least show you where the conflict is!

6304 to be competitive, or where there is some significant risk of this kind  
6305 of shenanigans, it's very easy to keep part, or all, of a repository private  
6306 among your collaborators until you are ready to share more widely. All  
6307 of the benefits we described still accrue. For an appropriately organized  
6308 and hosted project, often the only steps required to share materials, data,  
6309 and code are (1) to make the hosted repository public and (2) to link it  
6310 to an archival storage platform like the Open Science Framework.

### 6311 13.1.3 *File names*

6312 As Phil Karlton reportedly said,<sup>12</sup> “There are only two hard things in  
6313 Computer Science: cache invalidation and naming things.” What’s  
6314 true for computer science is true for research in general.<sup>13</sup> Naming  
6315 files is hard! Some very organized people survive on systems like  
6316 INFO-r1-draft-2020-07-13-js.docx, meaning “the INFO project  
6317 revision 1 draft of July 13th, 2020, with edits by JS.” But this kind of  
6318 system needs a lot of rules and discipline, and it requires everyone in a  
6319 project to buy in completely.

6320 On the other hand, if you are naming a file in a hierarchically organized  
6321 version control repository, the naming problem gets dramatically eas-  
6322 ier. All of a sudden, you have a context in which names make sense.  
6323 data.csv is a terrible name for a data file on its own. But the name

<sup>12</sup> <https://www.karlton.org/2017/12/naming-things-hard>

<sup>13</sup> We won’t talk about cache invalidation; that’s a more technical problem in computer science that is beyond the scope of this book.

6324 is actually perfectly informative—in the context of a project repository  
6325 with a README that states that there is only a single experiment, a  
6326 repository structure such that the file lives in a folder called `raw_data`,  
6327 and a commit history that indicates the file’s commit date and author.

6328 As this example shows, naming is hard *out of context*. So here’s our rule:  
6329 name a file with what it contains. Don’t use the name to convey the  
6330 context of who edited it, when, or where it should go in a project. That  
6331 is metadata that the platform should take care of.<sup>14</sup>

<sup>14</sup> The platform won’t take care of it if you email it to a collaborator—precisely why you should share access to the full *platform*, not just the *out-of-context* file!

## 6332 13.2 Data Management

6333 We’ve just discussed how to manage projects in general; in this section  
6334 we zoom in on datasets specifically. Data are often the most valuable  
6335 research product because they represent the evidence generated by our  
6336 research. We maximize the value of the evidence when other scientists  
6337 can reuse it for independent verification or generation of novel discov-  
6338 eries. Yet, lots of research data are not reusable, even when they are  
6339 shared. In chapter 3, we discussed Hardwicke et al.’s (2018) study of  
6340 *analytic* reproducibility. But before we were even able to try and re-  
6341 produce the analytic results, we had to look at the data. When we did  
6342 that, we found that only 64% of shared datasets were both complete and  
6343 understandable.

6344 How can you make sure that your data are managed so as to enable  
6345 effective sharing? We make four primary recommendations:

- 6346 1. save your raw data  
6347 2. document your data collection process  
6348 3. organize your raw data for later analysis  
6349 4. document your data using a codebook or other metadata

6350 Let's look at each in turn.

### 6351 13.2.1 Save your raw data

6352 Raw data take many forms. For many of us, the raw data are those re-  
6353 turned by the experimental software; for others, the raw data are videos  
6354 of the experiment being carried out. Regardless of the form of these  
6355 data, save them! They are often the only way to check issues in what-  
6356 ever processing pipeline brings these data from their initial state to the  
6357 form you analyze. They also can be invaluable for addressing critiques  
6358 or questions about your methods or results later in the process. If you  
6359 need to correct something about your raw data, *do not alter the original*  
6360 *files*. Make a copy, and make a note about how the copy differs from  
6361 the original.<sup>15</sup>

<sup>15</sup> Future you will thank present you for explaining why there are two copies of subject 19's data after you went back and corrected a typo.

6362 Raw data are often not anonymized—or even anonymizable.  
6363 Anonymizing them sometimes means altering them (e.g., in the  
6364 case of downloaded logs from a service that might include IDs or IP  
6365 addresses). Or in some cases, anonymization is difficult or impossible  
6366 without significant effort and loss of some value from the data, for  
6367 example, for video data or MRI data (Bischoff-Grethe et al. 2007).  
6368 Unless you have specific permission for broad distribution of these  
6369 identifiable data, the raw data may then need to be stored in a different  
6370 way. In these cases, we recommend saving your raw data in a separate  
6371 repository with the appropriate permissions. For example, in the  
6372 ManyBabies 1 study we described above, the public repository does not  
6373 contain the raw data contributed by participating labs, which the team  
6374 could not guarantee was anonymized; these data are instead stored in a  
6375 private repository.<sup>16</sup>

6376 You can use your repository’s README to describe what is and  
6377 is not shared. For example, a README might state, “We provide  
6378 anonymized versions of the files originally downloaded from Qualtrics”  
6379 or “Participants did not provide permission for public distribution  
6380 of raw video recordings, which are retained on a secure university  
6381 server.” Critically, if you share the derived tabular data, it should  
6382 still be possible to reproduce the analytic results in your paper, even  
6383 if checking the provenance of those numbers from the raw data is not

<sup>16</sup> The precise repository you use for this task is likely to vary by the kind of data that you’re trying to store and the local regulatory environment. For example, in the United States, to store de-anonymized data with certain fields requires a server that is certified for HIPAA (the relevant privacy law). Many—but by no means all—universities provide HIPAA-compliant cloud storage.

6384 possible for every reader.<sup>17</sup>

6385 One common practice is the use of participant identifiers to link spe-  
6386 cific experimental data—which, if they are responses on standardized  
6387 measures, rarely pose a significant identifiability risk—to demographic  
6388 data sheets that might include more sensitive and potentially identifi-  
6389 able data.<sup>18</sup> Depending on the nature of the analyses being reported,  
6390 the experimental data can then be shared with limited risk. Then a  
6391 selected set of demographic variables—for example, those that do not  
6392 increase privacy risks but are necessary for particular analyses—can be  
6393 distributed as a separate file and joined back into the data later.

<sup>17</sup> One way we organize the raw data in some of our paper is to have three different subfolders in the `data/` directory: `raw/`, for the original data; `processed/`, for the anonymized or otherwise preprocessed data; and `/scripts`, for the code that does the preprocessing. Since these folders are in a Git repository, we can then add `raw/*` to the `.gitignore` file, ensuring that they are never added to the public version of the repository even though they sit within our local file hierarchy in the appropriate place.

### 6394 13.2.2 Document your data collection process

6395 In order to understand the meaning of the raw data, it's helpful to share  
6396 as much as possible about the context in which they were collected.  
6397 This practice also helps communicate the experience that participants  
6398 had in your experiment. Documentation of this experience can take  
6399 many forms.

6400 If the experimental experience was a web-based questionnaire, archiv-  
6401 ing this experience can be as simple as downloading the questionnaire  
6402 source.<sup>19</sup> For more involved studies, it can be more difficult to recon-  
6403 struct what participants went through. This kind of situation is where

video data can shine (Gilmore and Adolph 2017). A video recording of a typical experimental session can provide a valuable tutorial for other experimenters—as well as good context for readers of your paper. This is doubly true if there is a substantial interactive element to your experimental experience, as is often the case for experiments with children. For example, in our ManyBabies case study, the project shared “walk-through” videos of experimental sessions<sup>20</sup> for many of the participating labs, creating a repository of standard experiences for infant development studies. If nothing else, a video of an experimental session can sometimes be a very nice archive of a particular context.<sup>21</sup>

Regardless of what specific documentation you keep, it’s critical to create some record linking your data to the documentation. For a questionnaire study, for example, this documentation might be as simple as a README that says that the data in the data/raw/ directory were collected on a particular date using the file named experiment1.qsf. This kind of “connective tissue” linking data to materials can be very important when you return to a project with questions. If you spot a potential error in your data, you will want to be able to examine the precise version of the materials that you used to gather those data in order to identify the source of the problem.

<sup>18</sup> A word about subject identifiers. These should be anonymous identifiers, like randomly generated numbers, that cannot be linked to participant identities (like date of birth) and are unique. You laugh, but one of us was in a lab where all the subject IDs were the date of test and the initials of the participant. These were neither unique nor anonymous. One common convention is to give your study a code-name and to number participants sequentially, so your first participant in a sequence of experiments on information processing might be INFO-1-01.

<sup>19</sup> If it’s in a proprietary format like a Qualtrics .QSF file, a good practice is to convert it to a simple plain text format as well so it can be opened and reused by folks who do not have access to Qualtrics (which may include future you!).

<sup>20</sup> <https://nyu.databrary.org/volume/>  
<sup>21</sup> 896 Videos of experimental sessions also are great demos to show in a presentation about your experiment, provided you have permission from the participant.

6424 13.2.3 Organize your data for later analysis: Spreadsheets

6425 Data come in many forms, but chances are that at some point during  
 6426 your project you will end up with a spreadsheet full of information.  
 6427 Well-organized spreadsheets can mean the difference between project  
 6428 success and failure! A wonderful article by Broman and Woo (2018) lays  
 6429 out principles of good spreadsheet design. We highlight some of their  
 6430 principles here (with our own, opinionated ordering):

6431 1. *Make it a rectangle.*<sup>22</sup> Nearly all data analysis software, like SPSS,  
 6432 Stata, Jamovi, and JASP (and many R packages), require data to be  
 6433 in a tabular format.<sup>23</sup> If you are used to analyzing data exclusively  
 6434 in a spreadsheet, this kind of tabular data isn't quite as readable,  
 6435 but readable formatting gets in the way of almost any analysis you  
 6436 want to do. Figure 13.4 gives some examples of nonrectangular  
 6437 spreadsheets. All of these will cause any analytic package to choke  
 6438 because of inconsistencies in how rows and columns are used!

A	B	C	D	E	F
1					
2	101	102	103	104	105
3 sex	Male	Female	Male	Male	Male
4					
5	101	102	103	104	105
6 glucose	134.1	120.8	124.8	83.1	105.2
7					
8	101	102	103	104	105
9 insulin	0.60	1.18	1.23	1.16	0.73

A	B	C	D	E	F	G
1	1MIN					
2		Normal			Mutant	
3 B6	146.6	138.6	155.6	166	179.3	186.9
4 BTBR	245.7	240				
5						
6 5MIN						
7		Normal			Mutant	
8 B6	333.6	353.6	408.8	450.6	474.4	423.8
9 BTBR	514.4	610.6	597.9	412.1	447.4	446.5

A	B	C	D	E	F
1	GTt date	GTt weight	time	glucose mg/dl	insulin ng/ml
2 Date	11/3/14			99.9	10 off curve
3 Days on Diet	126			309.3	0.295
4 Mouse #	43			15	284.1
5 sex	f			30	0.175
6 experiment	values		mean	60	99.9
7 control	0.186	0.191	1.081	0.49	0.122
8 treatment A	7.414	1.468	2.254	3.71	3.23
9 treatment B	9.811	9.259	11.296	10.12	1.05
10				120	217.9
11 fold change	values		mean	0	185.8
12 treatment A	15.26	3.82	4.64	7.64	6.65
13 treatment B	20.19	19.05	23.24	20.83	2.17

A	B	C	D	E	F	G
1	321	2/9/15	24.5	0	297.4	2.228
2				6	309.3	0.295
3				15	284.1	0.129
4				30	312	0.175
5				60	99.9	0.122
6				120	217.9	10 off curve
7				0	185.8	0.251
8	322	2/9/15	18.9	5	297.4	2.078
9				15	439	2.078
10				30	303.3	0.175
11				60	232.7	0.15
12				120	249.7	0.523
13				0	198.5	0.151
14	323	2/9/15	24.7	5	530.6	off curve 10
15						

<sup>22</sup> Think of your data like a well-ordered plate of sushi, neatly packed together without any gaps.

<sup>23</sup> Tabular data is a precursor to “tidy” data, which we describe in more detail in appendix D.

Figure 13.4 Examples of non-rectangular spreadsheet formats that are likely to cause problems in analysis. Adapted from Broman and Woo (2018).

6439 2. *Choose good names for your variables.* No one convention for name  
6440 formatting is best, but it's important to be consistent. We tend to  
6441 follow the tidyverse style guide<sup>24</sup> and use lowercase words sepa-  
6442 rated by underscores (\_). It's also helpful to give units where these  
6443 are available—for example, whether reaction times are in seconds  
6444 or milliseconds. Table 13.1 gives some examples of good and bad  
6445 variable names.

<sup>24</sup> <https://style.tidyverse.org>

Table 13.1  
Examples of good and bad variable names. Adapted from Broman and Woo (2018).

Good name	Good alternative	Avoid
subject_id	SubID	subject #
sex	female	M/F
rt_msec	reaction_time_ms	reaction time (millisec.)

6446 3. *Be consistent with your cell formatting.* Each column should have one  
6447 kind of thing in it. For example, if you have a column of numeri-  
6448 cal values, don't all of a sudden introduce text data like “missing”  
6449 into one of the cells. This kind of mixing of data types can cause  
6450 havoc down the road. Mixed or multiple entries also don't work,  
6451 so don't write “0 (missing)” as the value of a cell. Leaving cells  
6452 blank is also risky because it is ambiguous. Most software pack-  
6453 ages have a standard value for missing data (e.g., NA is what R  
6454 uses). If you are writing dates, please be sure to use the “global

standard” (ISO 8601), which is YYYY-MM-DD. Anything else

can be misinterpreted easily.<sup>25</sup>

4. *Decoration isn’t data.* Decorating your data with bold headings or highlighting may seem useful for humans, but it isn’t uniformly interpreted or even recognized by analysis software (e.g., reading an Excel spreadsheet into R will scrub all your beautiful highlighting and artistic fonts), so do not rely on it.

5. *Save data in plain text files.* The CSV (comma-delimited) file format is a common standard for data that is uniformly understood by most analysis software (it is an “interoperable” file format).<sup>26</sup> The advantage of CSVs is that they are not proprietary to Microsoft or another tech company and can be inspected in a text editor, but be careful: they do not preserve Excel formulas or formatting!

Given the points above, we recommend that you avoid analyzing your data in Excel. If it is necessary to analyze your data in a spreadsheet program, we urge you to save the raw data as a separate CSV and then create a distinct analysis spreadsheet so as to be sure to retain the raw data unaltered by your (or Excel’s) manipulations.

<sup>25</sup> Dates in Excel deserve special mention as a source of terribleness. Excel has an unfortunate habit of interpreting information that has nothing to do with dates as dates, destroying the original content in the process. Excel’s issue with dates has caused unending horror in the genetics literature, where gene names are automatically converted to dates, sometimes without the researchers noticing (Ziemann, Eren, and El-Osta 2016). In fact, some gene names have had to be changed in order to avoid this issue!

<sup>26</sup> Be aware of some interesting differences in how these files are output by European vs American versions of Microsoft Excel! You might find semicolons instead of commas in some datasets.

6474 13.2.4 *Organize your data for later analysis: Software*

6475 Many researchers do not create data by manually entering information  
6476 into a spreadsheet. Instead they receive data as the output from a web  
6477 platform, software package, or device. These tools typically provide re-  
6478 searchers limited control over the format of the resulting tabular data  
6479 export. Case in point is the survey platform Qualtrics, which—at least  
6480 at the moment—provides data with not one but two header rows, com-  
6481 plicating import into almost all analysis software!<sup>27</sup>

6482 That said, if your platform *does* allow you to control what comes out,  
6483 you can try to use the principles of good tabular data design outlined  
6484 above. For example, try to give your variables (e.g., questions in  
6485 Qualtrics) sensible names!

<sup>27</sup> The R package `qualtRics` (Ginn, O'Brien, and Silge 2024) can help with this.

 ACCIDENT REPORT

*Bad variable naming can lead to analytic errors!*

In our methods class, students often try to reproduce the original analyses from a published study before attempting to replicate the results in a new sample of participants. When Kengthsagn Louis looked at the code for the study she was interested in, she noticed that the variables in the analysis code were named horribly (presumably because they were output this way by the survey software). For example, one piece of Stata code looked like this:

```
gen recall1=.  
  
replace recall1=0 if Q21==1  
  
replace recall1=1 if Q21==3 | Q21==5 | Q21==6  
  
replace recall1=2 if Q21==2 | Q21==4 | Q21==7 | Q21==8  
  
replace recall1=0 if Q69==1  
  
replace recall1=1 if Q69==3 | Q69==5 | Q69==6  
  
replace recall1=2 if Q69==2 | Q69==4 | Q69==7 | Q69==8  
  
ta recall1
```

In the process of translating this code into R in order to reproduce the analyses, Kengthsagn and a course teaching assistant, Andrew Lampinen, noticed that some participant responses had been assigned to the wrong variables. Because the variable names were not human-readable, this error was almost impossible to detect. Since the problem affected some of the inferential conclusions of the article, the article's author—to their credit—issued an immediate correction ([M. B. Petersen 2019](#)).

The moral of the story: obscure variable names can hide existing errors and create opportunities for further error! Sometimes you can adjust these within your experimental software, avoiding the issue. If not, make sure to create a “key” and translate the names immediately, double checking after you are done.

6488 13.2.1 Document the format of your data

6489 Even the best-organized tabular data are not always easy to understand  
 6490 by other researchers, or even yourself, especially after some time has  
 6491 passed. For that reason, you should make a **codebook** (also known as  
 6492 a **data dictionary**) that explicitly documents what each variable is. Fig-  
 6493 ure 13.6 shows an example codebook for the trial-level data in the bot-  
 6494 tom of figure 13.5. Each row represents one variable in the associated  
 6495 dataset. Codebooks often describe what type of variable a column is  
 6496 (e.g., numeric, string), and what values can appear in that column. A  
 6497 human-readable explanation is often given as well, providing units (e.g.,  
 6498 “seconds”) and a translation of numeric codes (e.g., “test condition is  
 6499 coded as 1”) where relevant.

	A	B	C	D	E	F	G	H	I
1	lab	subid	method	RA	age_days	trial_order	session_error	session_error_type	notes
2	babylab_nijmegen	ba01_6-9	HPP	KM	245	1	noerror	NA	teeth may be painful
3	babylab_nijmegen	ba02_6-9	HPP	KM	206	4	noerror	NA	NA
4	babylab_nijmegen	ba03_6-9	HPP	KM	257	3	noerror	NA	NA
5	babylab_nijmegen	ba04_6-9	HPP	KM	245	2	error	baby cried	teeth may be painful
6	babylab_nijmegen	ba05_6-9	HPP	KM	288	2	noerror	NA	baby was sick 2 months ago

	A	B	C	D	E	F
1	lab	subid	trial_type	stimulus	trial_num	looking_time
2	babylab_nijmegen	ba01_6-9	training	train1	-2	18.02
3	babylab_nijmegen	ba01_6-9	training	train2	-1	9.05
4	babylab_nijmegen	ba01_6-9	IDS	IDS1	1	17.48
5	babylab_nijmegen	ba01_6-9	ADS	ADS1	2	5.51
6	babylab_nijmegen	ba01_6-9	IDS	IDS2	3	16.34
7	babylab_nijmegen	ba01_6-9	ADS	ADS2	4	13.9

	A	B	C	D
1	Variable Name	Type	Possible Values	Explanation
2	lab	string	<your lab ID>	your unique lab ID
3	subid	string	<participant ID codes>	unique (within lab) ID for the participant
4	trial_type	string	'IDS', 'ADS', and 'training'	stimulus type on this trial
5	stimulus	string	'IDS-x', 'ADS-x', 'training'	the actual sound file that was playing
6	trial_num	integer	-2, -1, 1-8	trial number, from 1 -- 8 (with -2 and -1 denoting training trials)
7	looking_time	double	range 0-20	looking time in seconds

Figure 13.5

Example participant (top) and trial (bottom) level data from the ManyBabies (2020) case study.

Figure 13.6

Codebook for trial-level data (see above) from the ManyBabies (2020) case study.

6500 Creating a codebook need not require a lot of work. Almost any docu-  
 6501 mentation is better than nothing! There are also several R packages that  
 6502 can automatically generate a codebook for you, for example codebook

6503 (Arslan 2019), **dataspice** (Boettiger et al. 2021), and **dataMaid** (A. H.  
6504 Petersen and Ekstrøm 2019). Adding a codebook can substantially in-  
6505 crease the reuse value of the data and prevent hours of frustration as  
6506 future you and others try to decode your variable names and assump-  
6507 tions.

### 6508 13.3 *Sharing Research Products*

6509 As we've been discussing throughout this chapter, if you've managed  
6510 your research products effectively, sharing them with others is a far less  
6511 daunting prospect, and usually just requires uploading them to an online  
6512 repository like the Open Science Framework. This section addresses  
6513 some potential limitations on sharing that you should bear in mind and  
6514 discusses where and how to share research products.

#### 6515 13.3.1 *What you can and can't share*

6516 We've been advocating that you share all of your research products, es-  
6517 pecially your data. In practice, however, **participant privacy** (as well as  
6518 a few other constraints) limits what you can share. Luckily, there are  
6519 some concrete steps you can take to make sure that you protect partici-  
6520 pants and comply with your obligations while still realizing the benefits  
6521 of data sharing.

6522 Unless they explicitly waive their rights, participants in psychology ex-  
6523 periments have the expectation of privacy—that is, no one should be  
6524 able to identify them from the data they have provided. Protecting par-  
6525 ticipant privacy is an important part of researchers' ethical responsibili-  
6526 ties (Ross, Iguchi, and Panicker 2018) and needs to be balanced against  
6527 the ethical imperatives to share (see chapter 4).<sup>28</sup>

6528 Furthermore, there are legal regulations that protect participants' data,  
6529 though these vary from country to country. In the US, the relevant reg-  
6530 ulation is **HIPAA**, the Health Insurance Portability and Accountability  
6531 Act, which limits disclosures of private health information (**PHI**). In the  
6532 European Union, the relevant regulation is the European **GDPR** (Gen-  
6533 eral Data Protection Regulation). It's beyond the scope of this book to  
6534 give a full treatment of these regulatory frameworks; you should con-  
6535 sult with your local ethics board regarding compliance, but here is the  
6536 way we have navigated this situation while still sharing data.

6537 Under both frameworks, **anonymization** (or equivalently de-  
6538 **identification**) of data is a key concept, such that data sharing is  
6539 generally just fine if the data meet the relevant standard. Under US  
6540 guidelines, researchers can follow the “safe harbor” standard<sup>29</sup> under  
6541 which data are considered to be anonymized if they do not contain  
6542 identifiers like names, telephone numbers, email addresses, social

<sup>28</sup> Meyer (2018) gives an excellent overview of how to navigate various legal and ethical issues around data sharing in the US context.

<sup>29</sup> As described on the relevant DHHS page (<https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>).

6543 security numbers, dates of birth, faces, and others. Thus, data that  
6544 only contain participant IDs and nothing from this list can typically be  
6545 shared without participant consent without a problem.<sup>30</sup>

6546 The EU's GDPR also allows fully anonymized data sharing, with one  
6547 big complication. Putting anonymous identifiers in a data file and re-  
6548 moving identifiable fields does not itself suffice for GDPR anonymiza-  
6549 tion if the data are still **in principle reidentifiable** because you have main-  
6550 tained documentation linking IDs to identifiable data like names or  
6551 email addresses. Only when the key linking identifiers to data has been  
6552 destroyed are the data truly de-identified according to this standard.

<sup>30</sup> US IRBs are a very decentralized bunch, and their interpretations often vary considerably. For reasons of liability or ethics, they may not allow data sharing even though it is permitted by US law. If you feel like arguing with an IRB that takes this kind of stand, you could mention that the DHHS rule actually doesn't consider de-identified data to be "human subjects" data at all, and thus the IRB may not have regulatory authority over it. We're not lawyers, and we're not sure if you'll succeed, but it could be worth a try.

## ❖ ACCIDENT REPORT

### *Really anonymous?*

When we first began teaching Psych 251, our experimental methods course at Stanford, one of the biggest contributions of the course was simply showing students how to do experiments online. Amazon's Mechanical Turk crowdsourcing service was relatively new, and our IRB did not have a good sense of what this service really was. We proposed that we would share data from the class and received approval for this practice. Our datasets were downloaded directly from Mechanical Turk and included participants' MTurk IDs (long alphanumeric strings that seemed completely anonymous). Several experiences caused us to reconsider this practice!

First, we discovered that MTurk IDs were in some cases linked to study participants' public Amazon "wish lists," which could both inadvertently provide information about the participant and also even potentially provide a basis for reidentification (in rare cases). This discovery led us to consult with our IRB and provide more explicit consent language in our class experiments, linking to instructions for making Amazon profiles private.

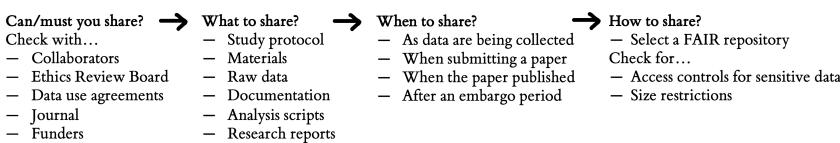
Then, a little later we received an irate email from an MTurk participant who had discovered their data on GitHub via a search for their MTurk ID. Although they were not identified in this dataset, it convinced us that at least some participants would not like this ID shared. After another consultation with the IRB, we apologized to this individual and removed their and others' IDs from our GitHub commit histories across that and other repositories. Prior to posting data, we now take care to anonymize IDs by creating a secret mapping between the IDs we post and the actual MTurk IDs.

6554

6555 De-identification is not always enough. As datasets get richer, **statistical**  
6556 **reidentification risks** go up substantially such that, with a little bit of out-  
6557 side information, data can be matched with a unique individual. These  
6558 risks are especially high with linguistic, physiological, and geospatial  
6559 data, but they can be present even for simple behavioral experiments.  
6560 In one influential demonstration, knowing a person's location on two  
6561 occasions was often enough to identify their data uniquely in a huge

6562 database of credit card transactions (De Montjoye et al. 2015).<sup>31</sup> Thus,  
 6563 simply removing fields from the data is a good starting point—but if you  
 6564 are collecting richer data about participants’ behavior you may need to  
 6565 consult an expert.

6566 Privacy issues are ubiquitous in data sharing, and almost every exper-  
 6567 imental research project will need to solve them before sharing data.  
 6568 For simple projects, often these are the only issues that preclude data  
 6569 sharing. However, in more complex projects, other concerns can arise.  
 6570 Funders may have specific mandates regarding where your data should  
 6571 be shared. Data use agreements or collaborator preferences may restrict  
 6572 where and when you can share. And certain data types require much  
 6573 more sensitivity since they are more consequential than, say, the reac-  
 6574 tion times on a Stroop task. We include here a set of questions to walk  
 6575 through to plan your sharing (figure 13.7). When in doubt, it’s often  
 6576 a good idea to consult with the relevant local authority—for example,  
 6577 your ethics board for ethical issues or your research management office  
 6578 for regulatory issues.



<sup>31</sup> For an example closer to home, many of the contributing labs in the Many-Babies project logged the date of test for each participant. This useful and seemingly innocuous piece of information is unlikely to identify any particular participant—but alongside a social media post about a lab visit or a dataset about travel records, it could easily reveal a particular participant’s identity.

Figure 13.7

A decision chart for thinking about sharing research products. Adapted from Klein et al. (2018).

6579    13.3.1 *Where and how to share: the FAIR principles*

6580    For shared research products<sup>32</sup> to be usable by others, they should meet  
6581    the FAIR standard by being findable, accessible, interoperable, and  
6582    reusable (Wilkinson et al. 2016).

- 6583    – **Findable** products are easily discoverable to both humans and  
6584    machines. That means linking to them in research reports  
6585    using unique persistent identifiers (e.g., a digital object identifier  
6586    [DOI])<sup>33</sup> and attaching them with metadata describing what they  
6587    are so they can be indexed by search engines.
- 6588    – **Accessibility** means that research products need to be preserved  
6589    across the long term and are retrievable via their standardized  
6590    identifier.
- 6591    – **Interoperability** means that the research products needs to be in a  
6592    format that people and machines (e.g., search engines and analysis  
6593    software) can understand.
- 6594    – **Reusable** means that the research products need to be well orga-  
6595    nized, documented, and licensed so that others know how to use  
6596    them.

<sup>32</sup> Most of this discussion is about data, because that's where the community has focused its efforts. That said, almost everything here applies to other research products as well!

<sup>33</sup> DOIs are those long URL-like things that are often used to link to papers. Turns out they can also be associated with datasets and other research products. Critically, they are guaranteed to work to find stuff, whereas standard web URLs often go stale after several years when people refactor their website. Most online repositories, like the Open Science Framework, will issue DOIs for the research products you store there.

6597    If you've followed the guidance in the rest of this chapter, then you will  
6598    already be well on your way to making your research products FAIR.

6599 There are a few final steps to consider. An important decision is where  
6600 you are going to share the research products. We recommend uploading  
6601 the files to a repository that's designed to support FAIR principles. Per-  
6602 sonal websites don't cut it, since these sites tend to go out of date and  
6603 disappear. There's also no easy way to find research products on per-  
6604 sonal sites unless you know who created them. GitHub, though it's a  
6605 great platform for collaboration, isn't a FAIR repository—for one thing,  
6606 products there don't necessarily have DOIs<sup>34</sup>—and there are no archival  
6607 guarantees on files that are shared there. Perhaps surprisingly for some  
6608 researchers, journal supplementary materials are also not a great place to  
6609 put research products. Often supplementary materials are assigned no  
6610 unique DOI or metadata, have limited supported formats, and have no  
6611 persistence guarantees (Evangelou, Trikalinos, and Ioannidis 2005).

6612 Fortunately, there are many repositories that help you conform to FAIR  
6613 standards. Zenodo, Figshare, the Open Science Framework (OSF), and  
6614 the various Dataverse sites are designed for this purpose, though there  
6615 are many other domain-specific repositories that are particularly rele-  
6616 vant for different research fields. We often use the OSF as it makes it  
6617 easy to share all research products connected to a project in one place.  
6618 Open Science Framework is FAIR compatible and allows users to assign  
6619 DOIs to their data and provide appropriate metadata.

<sup>34</sup> You can get a DOI for GitHub soft-  
ware through a partnership with Zenodo  
(<https://zenodo.org>), a FAIR-compliant  
repository.

6620 We recommend you attach a license to your research products. Academic culture is (usually) unburdened by discussion of intellectual property and legal rights and instead relies on scholarly norms about citation  
6621 and attribution. The basic expectation is that if you rely on someone else's research, you explicitly acknowledge the relevant journal article  
6622 through a citation. Although norms are still evolving, using research products created by others generally adheres to the same scholarly principle.  
6623 Research products can also be useful in nonacademic contexts, however. Perhaps you created software that a company would like  
6624 to use. Maybe a pediatrician would like to use a research instrument  
6625 you've been working on to assess their patients. These applications (and many other reuses of the data) require a legal license. In practice, there  
6626 are a number of simple, open-source licenses that permit reuse. We  
6627 tend to favor Creative Commons<sup>35</sup> licenses, which come in a variety of  
6628 flavors such as CC0<sup>36</sup> (which allows all reuse), CC-BY<sup>37</sup> (which allows  
6629 reuse as long as there is attribution), and CC-BY-NC<sup>38</sup> (which only  
6630 allows attributed, noncommercial reuse).<sup>39</sup> Regardless of what license  
6631 you choose, having a license means that your products won't be in a  
6632 "not sure what I'm allowed to do with this" limbo for others who are  
6633 interested in reusing them.

6640 As we have discussed, you may want to consider storing your work in a public repository from the outset of the project. If you are using GitHub

<sup>35</sup> <https://creativecommons.org>

<sup>36</sup> <https://creativecommons.org/share-your-work/public-domain/cc0>

<sup>37</sup> <https://creativecommons.org/licenses/by/4.0>

<sup>38</sup> <https://creativecommons.org/licenses/by/4.0>

<sup>39</sup> Klein et al. (2018) recommend the CC0 license, which puts no limits on what can be done with your data. At first glance, it may seem like a license that requires attribution is useful. But academic norms, rather than the threat of litigation, lead to good citation practices.

In addition, more restrictive licenses can mean that some legitimate uses of your data or research can be blocked.

6642 to manage your project, you can link the Git repository to the Open  
6643 Science Framework so it automatically syncs. This provides a valuable  
6644 incentive to organize your work properly throughout your project and  
6645 makes sharing super easy, because you've already done it! On the other  
6646 hand, this way of working can feel exposed for some researchers, and it  
6647 does carry some risks, however small, of "scooping" or preemption by  
6648 other groups working in the same space. Fortunately you can set up the  
6649 same Git-OSF workflow and keep it private until you're ready to make  
6650 it public later on.

6651 The next stage at which you should consider sharing your research prod-  
6652 ucts is when you submit your study to a journal. If you're still hesitant  
6653 to make the project entirely public, many repositories (including OSF)  
6654 will allow you to create special links that facilitate limited access to, for  
6655 example, reviewers and editors. In general, the earlier you share your  
6656 research products the better because there are more opportunities for  
6657 others to learn from, build on, and verify your research.<sup>40</sup> But if neither  
6658 of these options seems appealing, please do share your research products  
6659 once your paper is accepted. Doing so will increase the value (and the  
6660 impact) of your publication.

<sup>40</sup> If there are errors in our work, we'd certainly love to hear about it *before* the article is published in a journal rather than after!

### 6661 13.4 Chapter summary

6662 All of the hard work you put into your experiments—not to mention  
6663 the contributions of your participants—can be undermined by bad data  
6664 and project management. As our accident reports and case study show,  
6665 bad organizational practices can at a minimum cause huge headaches.  
6666 Sometimes the consequences can be even worse. On the flip side, start-  
6667 ing with a firm organizational foundation sets your experiment up for  
6668 success. These practices also make it easier to share all of the products  
6669 of your research, not just your findings. Such sharing is useful both for  
6670 individual researchers and for the field as a whole.



#### DISCUSSION QUESTIONS

1. Find an Open Science Framework repository that corresponds to a published paper. What is their strategy for documenting what is shared? How easy is it to figure out where everything is and if the data and materials sharing is complete?
2. Open up the US Department of Health and Human Services “safe harbor” standards (<https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>) and navigate to the section called “The De-identification Standard.” Go through the list of identifiers that must be removed. Are there any on this list that you would need to include in your dataset in order to conduct your own research? Can you think of any others that do not

fall on this list?

6672

## READINGS

- A more in-depth tutorial on various aspects of scientific openness:  
Klein, Olivier, Tom E Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C Frank (2018). “A Practical Guide for Transparency in Psychological Science.” *Collabra: Psychology* 4 (1): 20. <https://doi.org/10.1525/collabra.158>

6673

6674  
V

6675  
**REPORTING**

6676 *References*

- Arslan, Ruben C. 2019. “How to Automatically Document Data with the Codebook Package to Facilitate Data Reuse.” *Advances in Methods and Practices in Psychological Science* 2 (2): 169–187. <https://doi.org/10.1177/2515245919838783>.
- Bischoff-Grethe, Amanda, I. Burak Ozyurt, Evelina Busa, Brian T. Quinn, Christine Fennema-Notestine, Camellia P. Clark, Shaunna Morris, et al. 2007. “A Technique for the Deidentification of Structural Brain MR Images.” *Human Brain Mapping* 28 (9): 892–903.
- Blischak, John D., Emily R. Davenport, and Greg Wilson. 2016. “A Quick Introduction to Version Control with Git and GitHub.” *PLoS Computational Biology* 12 (1): e1004668.
- Boettiger, Carl, Scott Chamberlain, Auriel Fournier, Kelly Hondula, Anna Krystalli, Bryce Mecum, Maëlle Salmon, Kate Webbink, and Kara Woo. 2021. *dataspice: Create Lightweight Schema.org Descriptions of Data*. <https://CRAN.R-project.org/package=dataspice>.
- Broman, Karl W., and Kara H. Woo. 2018. “Data Organization in Spreadsheets.” *The American Statistician* 72 (1): 2–10.
- Byers-Heinlein, Krista, Christina Bergmann, Catherine Davies, Michael C. Frank, J. Kiley Hamlin, Melissa Kline, Jonathan F. Kominsky, et al. 2020. “Building a Collaborative Psychological Science: Lessons Learned from ManyBabies 1.” *Canadian Psychology/Psychologie Canadienne* 61 (4): 349–363.
- De Montjoye, Yves-Alexandre, Laura Radaelli, Vivek Kumar Singh, et al. 2015. “Unique in the Shopping Mall: On the Reidentifiability of Credit

- Card Metadata.” *Science* 347 (6221): 536–539.
- Evangelou, Evangelos, Thomas A. Trikalinos, and John P. A. Ioannidis. 2005. “Unavailability of Online Supplementary Scientific Information from Articles Published in Major Journals.” *The FASEB Journal* 19 (14): 1943–1944.
- Gilmore, Rick O., and Karen E. Adolph. 2017. “Video Can Make Behavioural Science More Reproducible.” *Nature Human Behaviour* 1, 0128 (2017). <https://doi.org/10.1038/s41562-017-0128>.
- Ginn, Jasper, Joseph O’Brien, and Julia Silge. 2024. *qualRics: Download “Qualtrics” Survey Data.* <https://CRAN.R-project.org/package=qualRics>.
- Hardwicke, Tom E., Manuel Bohn, Kyle MacDonald, Emily Hembacher, Michèle B. Nuijten, Benjamin N. Peloquin, Benjamin E. deMayo, Bria Long, Erica J. Yoon, and Michael C. Frank. 2021. “Analytic Reproducibility in Articles Receiving Open Data Badges at the Journal Psychological Science: An Observational Study.” *Royal Society Open Science* 8 (1): 201494. <https://doi.org/10.1098/rsos.201494>.
- Hardwicke, Tom E., and John P. A. Ioannidis. 2018. “Populating the Data Ark: An Attempt to Retrieve, Preserve, and Liberate Data from the Most Highly-Cited Psychology and Psychiatry Articles.” *PLOS ONE* 13 (8): e0201856. <https://doi.org/10.1371/journal.pone.0201856>.
- Hardwicke, Tom E., Maya B. Mathur, Kyle Earl MacDonald, Gustav Nilsonne, George Christopher Banks, Mallory Kidwell, Alicia Hofelich Mohr, et al. 2018. “Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition.” *Royal Society Open Science* 5. <https://doi.org/10.1098/rsos.180448>.

Houtkoop, Bobby Lee, Chris Chambers, Malcolm Macleod, Dorothy V. M.

Bishop, Thomas E. Nichols, and Eric-Jan Wagenmakers. 2018. “Data Sharing in Psychology: A Survey on Barriers and Preconditions.” *Advances in Methods and Practices in Psychological Science* 1 (1): 70–85. <https://doi.org/10.1177/2515245917751886>.

King, Gary, and Stuart Shieber. 2013. “Office Hours: Open Access.” YouTube. <https://www.youtube.com/watch?v=jD6CcFxRelY/>.

Klein, Olivier, Tom E. Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C. Frank. 2018. “A Practical Guide for Transparency in Psychological Science.” *Collabra: Psychology* 4 (1): 20. <https://doi.org/10.1525/collabra.158>.

Meyer, Michelle N. 2018. “Practical Tips for Ethical Data Sharing.” *Advances in Methods and Practices in Psychological Science* 1 (1): 131–144.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. “A Manifesto for Reproducible Science.” *Nature Human Behaviour* 1 (1): 1–9. <https://doi.org/10.1038/s41562-016-0021>.

Nosek, Brian A., George Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck, et al. 2015. “Promoting an Open Research Culture.” *Science* 348 (6242): 1422–1425. <https://doi.org/10.1126/science.aab2374>.

Petersen, Anne Helby, and Claus Thorn Ekstrøm. 2019. “dataMaid: Your Assistant for Documenting Supervised Data Quality Screening in R.” *Journal*

- of Statistical Software* 90 (6): 1–38. <https://doi.org/10.18637/jss.v090.i06>.
- Petersen, Michael Bang. 2019. “Corrigendum: Healthy Out-Group Members Are Represented Psychologically as Infected in-Group Members.” *Psychological Science* 30 (12): 1792–1794. <https://doi.org/10.1177/0956797619887750>.
- Piwowar, Heather A., and Todd J. Vision. 2013. “Data Reuse and the Open Data Citation Advantage.” *PeerJ* 1:e175.
- Ross, Michael W., Martin Y. Iguchi, and Sangeeta Panicker. 2018. “Ethical Aspects of Data Sharing and Research Participant Protections.” *American Psychologist* 73 (2): 138–145.
- Rouder, Jeffrey N. 2015. “The What, Why, and How of Born-Open Data.” *Behavior Research Methods* 48 (3): 1062–1069. <https://doi.org/10.3758/s13428-015-0630-z>.
- Simonsohn, Uri. 2013. “Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone.” *Psychological Science* 24 (10): 1875–1888. <https://doi.org/10.1177/0956797613480366>.
- Tenopir, Carol, Natalie M. Rice, Suzie Allard, Lynn Baird, Josh Borycz, Lisa Christian, Bruce Grant, Robert Olendorf, and Robert J. Sandusky. 2020. “Data Sharing, Management, Use, and Reuse: Practices and Perceptions of Scientists Worldwide.” Edited by Sergi Lozano. *PLOS ONE* 15 (3): e0229003. <https://doi.org/10.1371/journal.pone.0229003>.
- The ManyBabies Consortium, Michael C. Frank, Katherine Jane Alcock, Natalia Arias-Trejo, Gisa Aschersleben, Dare Baldwin, Stéphanie Barbu, et al. 2020. “Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference.” *Advances in Methods and*

*Practices in Psychological Science* 3 (1): 24–52. <https://doi.org/10.1177/2515245919900809>.

Voytek, Bradley. 2016. “The Virtuous Cycle of a Data Ecosystem.” *PLOS Computational Biology* 12 (8): e1005037. <https://doi.org/10.1371/journal.pcbi.1005037>.

Wilkinson, Mark D., Michel Dumontier, I. J. S. brand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>.

Ziemann, Mark, Yotam Eren, and Assam El-Osta. 2016. “Gene Name Errors  
6681 Are Widespread in the Scientific Literature.” *Genome Biology* 17 (1): 1–3.

## 6682 14 WRITING



### LEARNING GOALS

- Write clearly by being concise, using structure, and adjusting to your audience
- Write reproducibly by interleaving writing and analysis code
- Write responsibly by acknowledging limitations, correcting errors, and calibrating your conclusions

6683

6684 You've designed and run your experiment, and you have even analyzed  
6685 your data. This final section of *Experimentology* discusses reporting your  
6686 results. We begin by thinking through how to write clearly, repro-  
6687 ducibility, and responsibly (this chapter); then we turn to the ques-  
6688 tion of designing informative and pretty data visualizations (chapter 15).

6689 Our final chapter in the section introduces meta-analysis as a tool for  
6690 research synthesis, allowing us to contextualize research results. These  
6691 chapters focus on themes of TRANSPARENCY as well as (especially for  
6692 meta-analysis) BIAS REDUCTION and MEASUREMENT PRECISION.

6693 All of the effort you put into designing and running an effective ex-  
6694 periment may be wasted if you cannot clearly communicate what you  
6695 did. Writing is a powerful tool—though you contribute to the conver-  
6696 sation only once, it enables you to speak to a potentially infinite num-  
6697 ber of readers. So it's important to get it right! In this chapter, we'll  
6698 provide some guidance on how to write scientific papers—the primary  
6699 method for reporting on experiments—clearly, reproducibly, and re-  
6700 sponsibly.<sup>1</sup>

### 6701 *14.1 Writing clearly*

6702 What is the purpose of writing? “Telepathy, of course,” says Stephen  
6703 King ([King 2000](#)). The goal of writing is to transfer information from  
6704 your mind to the reader’s as effectively as possible. Unfortunately, for  
6705 most of us, writing clearly does not come naturally; it is a craft we need  
6706 to work at.

6707 One of the most effective ways to learn to write clearly is to read and to  
6708 imitate the writing you admire. Many scientific articles are not clearly  
6709 written, so you will need to be selective in which models you imitate.  
6710 Fortunately, as a reader, you will know good writing when you see it—  
6711 you will feel like the writer is sending ideas directly from their mind to  
6712 yours. When you come across writing like that, try to find more work

<sup>1</sup> Clarity of communication was a founding principle of modern science. Early proto-scientists conducting alchemical experiments often made their work deliberately obscure—even writing in cryptic codes—so that others could not discover the “powerful secrets of nature.” Pioneers of scientific methodology, like Francis Bacon and Robert Boyle, pushed instead for transparency and clarity. Notoriously, Isaac Newton (originally an alchemist and later a scientist), continued to write in a deliberately obscure fashion in order to “protect” his work ([Heard 2016](#)).

6713 by the same author. The more good scientific writing you are exposed  
6714 to, the more you will develop a sense of what works and what does not.  
6715 You may pick up bad habits as well as good ones (we sure have!), but  
6716 over time, your writing will improve if you make a conscious effort to  
6717 weed out the bad, and keep the good.

6718 There are no strict rules of clear writing, but there are some generally  
6719 accepted conventions that we will share with you here, drawing from  
6720 both general style guides and those specific to scientific writing ([Zinsser](#)  
6721 [2006; Heard 2016; Gernsbacher 2018; Savage and Yeh 2019](#)).

#### 6722 14.1.1 *The structure of a scientific paper*

6723 A scientific paper is not a novel. Rather than reading from beginning  
6724 to end, readers typically jump between sections to extract information  
6725 efficiently ([Doumont 2009](#)). This “random access” is possible because  
6726 research articles typically follow the same conventional structure (see  
6727 figure 14.1). The main body of the article includes four main sections:  
6728 introduction, methods, results, and discussion (IMRaD).<sup>2</sup> This structure  
6729 has a narrative logic: What’s the knowledge gap? (introduction); how  
6730 did you address it? (methods); what did you find? (results); what do the  
6731 results mean? (discussion).

<sup>2</sup> In the old old days, there were few conventions—scientists would share their latest findings by writing letters to each other. But as the number of scientists and studies increased, this approach became unsustainable. The IMRaD structure gained traction in the 1800s and became dominant in the mid-1900s as scientific productivity rapidly expanded in the post-war era. We think IMRaD style articles are a big improvement, even if it is nice to receive a letter every now and again.

6732 Structure helps writers as well as readers. Try starting the writing pro-  
 6733 cess with section headings as a structure, then flesh it out, layer by layer.  
 6734 In each section, start by making a list of the key points you want to con-  
 6735 vey, each representing the first sentence of a new paragraph. Then add  
 6736 the content of each paragraph, and you'll be well on your way to having  
 6737 a full first draft of your article.

6738 Imagine that the breadth of focus in the body of your article has an  
 6739 “hourglass” structure (figure 14.1). The start of the introduction should  
 6740 have a broad focus, providing the reader with the general context of  
 6741 your study. From there, the focus of the introduction should get in-  
 6742 creasingly narrow until you are describing the specific knowledge gap  
 6743 or problem you will address and (briefly) how you are going to address  
 6744 it. The methods and results sections are at the center of the hourglass  
 6745 because they are tightly focused on your study alone. In the discussion  
 6746 section, the focus shifts in the opposite direction, from narrow to broad.  
 6747 Begin by summarizing the results of your study, discuss limitations, then  
 6748 integrate the findings with existing literature and describe practical and  
 6749 theoretical implications.

6750 Research articles are often packed with complex information; it is easy  
 6751 for readers to get lost. A “cross-reference” is a helpful signpost that tells  
 6752 readers where they can find relevant additional information without dis-

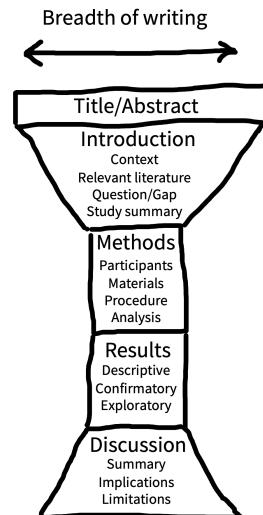


Figure 14.1  
 Conventional structure of a research article. The main body of the article consists of introduction, methods, results, and discussion (IMRaD) sections.

6753 rupting the flow of your writing. For example, you can refer the reader  
6754 to data visualizations by cross-referencing to figures or tables (e.g., “see  
6755 Figure 1”), or additional methodological information in the supplemen-  
6756 tary information (e.g., “see Supplementary Information A”).

6757 One useful trick for structuring complex arguments is to cross-  
6758 reference your research aims/hypotheses with your results. For  
6759 example, you could introduce numbered hypotheses in the intro-  
6760 duction of an article and then refer to them directly when reporting  
6761 the relevant analyses and results. These cross-references can serve to  
6762 remind readers how different results or analyses relate back to your  
6763 research goals.

#### 6764 *14.1.2 Paragraphs, sentences, and words*

6765 Writing an article is like drawing a human form. If you begin by sketch-  
6766 ing the clothes, you risk adding beautiful textures onto an impossible  
6767 shape. Instead, you have to start by understanding the underlying skele-  
6768 ton, and then gradually adding layers until you can visualize how cloth  
6769 hangs on the body. The structure of an article is the “skeleton” and  
6770 the paragraphs and sentences are the “flesh.” Only start thinking about  
6771 paragraphs and sentences once you have a solid outline in place.

6772 Ideally, each paragraph should correspond to a single point in the arti-  
6773 cle's outline, with the specifics necessary to convince the reader embed-  
6774 ded within. "P-E-E-L" (point - explain - evidence - link) is a useful  
6775 paragraph structure, particularly in the introduction and discussion sec-  
6776 tions. First, state the paragraph's message succinctly in the first sentence  
6777 (P). The core of the paragraph is dedicated to further explaining the  
6778 point and providing evidence (E-E; you can also include a third "E"—  
6779 an example). At the end of the paragraph, take a couple of sentences  
6780 to remind the reader of your point and set up a link to the next para-  
6781 graph.

6782 Since each sentence in a paragraph has a purpose, you can compose and  
6783 edit the sentence by asking how its form serves that purpose. For ex-  
6784 ample, short sentences are great for making strong initial points. On  
6785 the other hand, if you only use short sentences, your writing may come  
6786 across as monotonous and robotic. Try varying sentence lengths to give  
6787 your writing a more natural rhythm. Just avoid cramming too much in-  
6788 formation into the same sentence; very long sentences can be confusing  
6789 and difficult to process.

6790 You can also use sentence structure as a scaffold to support the reader's  
6791 thinking. Start sentences with something the reader already knows. For  
6792 example, rather than writing "We performed a between-subjects *t*-test

6793 comparing performance in the experimental and control groups to ad-  
6794 dress the cognitive dissonance hypothesis,” write “To address the cog-  
6795 nitive dissonance hypothesis, we compared performance in the experi-  
6796 mental group and control group using a between-subjects t-test.”

6797 Human readers are good at processing narratives about people. Yet, of-  
6798 ten scientists compromise the research narrative by removing themselves  
6799 from the process, sometimes even using awkward grammatical construc-  
6800 tions to do so. For example, scientists sometimes write “The data were  
6801 analysed” or, worse, “An analysis of the data was carried out.” Many of  
6802 us were taught to write sentences like these, but it’s much clearer to say  
6803 “We analyzed the data.”

6804 Similarly, many of us tend to hide our views with frames and caveats:  
6805 “[It is believed that/Research indicates that/Studies show that] money  
6806 leads to increased happiness (Frog & Toad, 1963).” If you truly do be-  
6807 lieve that money causes happiness, simply assert it—with a citation if  
6808 necessary. Save caveats for cases where *someone* believes that money  
6809 causes happiness, but it’s *not* you. Emphasize uncertainty where you in  
6810 fact feel that uncertainty is warranted, and readers will take your doubts  
6811 more seriously.

## 6812 14.2 Advice

6813 Scientific writing has a reputation for being dry, dull, and soulless.

6814 While it's true that writing research articles is more constrained than

6815 writing fiction, there are still ways to surprise and entertain your reader

6816 with metaphor, alliteration, and even humor. As long as your writing

6817 is clear and accurate, we see no reason why you cannot also make

6818 it enjoyable. Enjoyable articles are easier to read and more fun to

6819 write.<sup>3</sup>

6820 Here are a few more pieces of advice about expressing yourself

6821 clearly.

6822 **Be explicit.** Avoid vagueness and ambiguity. The more you leave the

6823 meaning of your writing to your reader's imagination, the greater the

6824 danger that different readers will imagine different things! So be direct

6825 and specific.

6826 **Be concise.** Maximize the signal-to-noise ratio in your writing by omit-

6827 ting needless words and removing clutter (Zinsser 2006). For example,

6828 say *we investigated* rather than *we performed an investigation of* and say *if*

6829 rather than *in the event that*. Don't try to convey everything you know

6830 about a topic—a research report is not an essay. Include only what you

6831 need to achieve the purpose of the article and exclude everything else.

<sup>3</sup> One of our favorite examples of an enjoyable article is Cutler (1994), a delightful piece that uses the form of the article to make a point about human language processing. Read it: you'll see!

6832 **Be concrete.** Concrete examples make abstract ideas easier to grasp. But  
6833 some ideas are just hard to express in prose, and diagrams can be very  
6834 helpful in these cases. For example, it may be clearer to illustrate a com-  
6835 plex series of exclusion criteria using a flow chart rather than text. You  
6836 can even use photos, videos, and screenshots to illustrate experimental  
6837 tasks (Heycke and Spitzer 2019).

6838 **Be consistent.** Referring to the same concept using different words can  
6839 be confusing because it may not be clear if you are referring to a different  
6840 concept or just using a synonym. For example, in everyday conversation,  
6841 “replication” and “reproducibility” may sound like two different ways  
6842 to refer to the same thing, but in scientific writing, these two concepts  
6843 have different technical definitions, so we should not use them inter-  
6844 changeably. Define each technical term once and then use the same  
6845 term throughout the manuscript.

6846 **Adjust to your audience.** Most of us adjust our conversation style de-  
6847 pending on who we’re talking to; the same principle applies to good  
6848 writing. Knowing your audience is more difficult with writing, because  
6849 we cannot see the reader’s reactions and adjust accordingly. Neverthe-  
6850 less, we can make some educated guesses about who our readers might  
6851 be. For example, if you are writing an introductory review article, you  
6852 may need to pay more attention to explaining technical terms than if

6853 you are writing a research article for a specialty journal.

6854 **Check your understanding.** Unclear writing can be a symptom of un-  
6855 clear thinking. If an idea doesn't make sense in your head, how will it  
6856 ever make sense on the page? In fact, trying to communicate something  
6857 in writing is an excellent way to probe your understanding and expose  
6858 logical gaps in your arguments. So, if you are finding it difficult to write  
6859 clearly, stop and ask yourself, *do I know what I want to say?* If the problem  
6860 is unclear thinking, then it might be worth talking out the ideas with a  
6861 colleague or advisor before you try to write them down.

6862 **Use acronyms sparingly.** It's tempting to replace lengthy terminology  
6863 with short acronyms—why say “cognitive dissonance theory” when  
6864 you can say “CDT”? Unfortunately, acronyms can increase the reader’s  
6865 cognitive burden and cause misunderstandings.<sup>4</sup> For example, if you  
6866 shorten “odds ratio” to “OR,” the reader has to take the extra step  
6867 of translating “OR” back to “odds ratio” every time they encounter  
6868 it. The problem multiplies as you introduce more acronyms into your  
6869 article. Worse, for some readers, “OR” tends to mean “operating  
6870 room,” not “odds ratio.” Acronyms can be useful, but usually only  
6871 when they are widely used and understood.

<sup>4</sup> Barnett and Doubleday (2020) found that acronyms are widely used in research articles and argued that they undermine clear communication. Here is one example of text Barnett and Doubleday extracted from a 2019 publication to illustrate the point: “Applying PROBAST showed that ADO, B-AE-D, B-AE-D-C, extended ADO, updated ADO, updated BODE, and a model developed by Bertens et al. were derived in studies assessed as being at low risk of bias.”

### 6872 14.2.1 *Drafting and revision*

6873 The clearest and most effortless-seeming scientific writing has proba-  
6874 bly gone through extensive revision to appear that way. It can sur-  
6875 prise many students to know the amount of revision that has gone into  
6876 many “breezy” articles. For example, Tversky and Kahneman repeat-  
6877 edly drafted and redrafted each word of their famous (and highly read-  
6878 able) articles on judgment and decision-making, hunched over the type-  
6879 writer together (Lewis 2016).

6880 Think of the article you are writing as a garden. Your first draft may  
6881 be an unruly mess of intertwined fronds and branches. Several rounds  
6882 of pruning and sculpting will be needed before your writing reaches its  
6883 most effective form. You’ll be amazed how often you find words you  
6884 can omit or elaborate sentences you can simplify.

6885 It can be difficult to judge if your own writing has achieved its tele-  
6886 pathic goal, especially after several rounds of revision. Try to get feed-  
6887 back from somebody in your target audience. Their comments—even  
6888 if not wholly positive—will give you a good sense of how much of your  
6889 argument they understood (and agreed with).<sup>5</sup>

<sup>5</sup> Seek out people who are willing to tell you that your writing is not good! They may not make you feel good, but they will help you improve.

### 6890 14.3 Writing reproducibly

6891 Many research results are not reproducible—that is, the numbers and  
6892 graphs that they report can't be recreated by repeating the original  
6893 analyses—even on the original data. As we discussed in chapter 3, a  
6894 lack of reproducibility is a big problem for the scientific literature; if  
6895 you can't trust the numbers in the articles you read, it's much harder to  
6896 build on the literature.

6897 Fortunately, there are number of tools and techniques available that you  
6898 can use to write fully reproducible research reports. The basic idea is to  
6899 create an unbroken chain that links every single part of the data analysis  
6900 pipeline, from the raw data through to the final numbers reported in  
6901 your research article. This linkage enables you—and hopefully others as  
6902 well—to trace the provenance of every number and recreate (reproduce)  
6903 it from scratch.

#### 6904 14.3.1 Why write reproducible reports?

6905 There are (at least) three reasons to write reproducible reports. First,  
6906 data analysis is an error-prone activity. Without safeguards in place, it  
6907 can be easy to accidentally overwrite data, mislabel experimental con-  
6908 ditions, or copy and paste the wrong statistics. As we discussed in chap-  
6909 ter 3, one study found that nearly half of a sample of psychology papers

6910 contained obvious statistical reporting errors (Nuijten et al. 2016). You  
6911 can reduce opportunities for error by adopting a reproducible analysis  
6912 workflow that avoids error-prone manual actions, like copying and past-  
6913 ing.

6914 Second, technical information about data analysis can be difficult to  
6915 communicate in writing. Prose is often ambiguous, and authors can in-  
6916 advertently leave out important details (Hardwicke et al. 2018). By con-  
6917 trast, a reproducible workflow documents the entire analysis pipeline  
6918 from raw data to research report exactly as it was implemented, describ-  
6919 ing the origin of any reported values and allowing readers to assess, ver-  
6920 ify, and repeat the analysis process.

6921 Finally, reproducible workflows are typically more efficient workflows.  
6922 For example, you may realize you forgot to perform data exclusions and  
6923 need to rerun the analysis. You may produce a graph and then decide  
6924 you'd prefer a different color scheme. Or perhaps you want to output  
6925 the same results table in a PDF document and in a PowerPoint slide. In  
6926 a reproducible workflow, all of the analysis steps are scripted and can  
6927 be easily rerun at the click of a button. You (and others) can also reuse  
6928 parts of your code in other projects, rather than having to write from  
6929 scratch.

### 6930 14.3.2 Principles of reproducible writing

6931 Below we outline some general principles of reproducible writing.  
6932 These can be put in practice in a number of different software ecosys-  
6933 tems. We recommend R Markdown and its successor, Quarto, which  
6934 are ways of writing data analysis code in R so that it compiles into  
6935 spiffy documents or even websites. (This book was written in Quarto.)  
6936 Chapter C gives an introduction to the nuts and bolts of using these  
6937 tools to create scientific papers.

6938 – **Never break the chain.** Every part of the analysis pipeline—from  
6939 raw data<sup>6</sup> to final product—should be present in the project  
6940 repository. By consulting the repository documentation, a reader  
6941 should be able to follow the steps to go from the raw data to the  
6942 final manuscript, including tables and figures.

6943 – **Script everything.** Try to ensure that each step of the analysis  
6944 pipeline is executed by computer code rather than manual ac-  
6945 tions such as copying and pasting or directly editing spreadsheets.

6946 This practice ensures that every step is documented via executable  
6947 code rather than ambiguous description, ensuring that it can be  
6948 reproduced. Imagine, for example, that you decided to recode  
6949 a variable in your dataset. You could use the “find and replace”  
6950 function in Excel, but this action would not be documented—you

<sup>6</sup> Modulo the privacy concerns discussed in chapter 13, of course.

6951 might even forget that you did it! A better option would be to  
6952 write an R script. While a scripted pipeline can be a pain to set  
6953 up the first time, by the third time you rerun it, it will save you  
6954 time.

6955 – **Use literate programming.** The meaning of a chunk of computer  
6956 code is not always obvious to another user, especially if they’re  
6957 not an expert. Indeed, we frequently look at our own code and  
6958 scratch our heads, wondering what on earth it’s doing. To avoid  
6959 this problem, try to structure your code around plain language  
6960 comments that explain what it should be doing, a technique  
6961 known as “literate programming” ([Knuth 1992](#)).

6962 – **Use defensive programming.** Errors can still occur in scripted  
6963 analyses. Defensive programming is a series of strategies to help  
6964 anticipate, detect, and avoid errors in advance. A typical defensive  
6965 programming tool is the inclusion of **tests** in your code, snippets  
6966 that check if the code or data meet some assumptions. For exam-  
6967 ple, you might test if a variable storing reaction times has taken on  
6968 values below zero (which should be impossible). If the test passes,  
6969 the analysis pipeline continues; if the test fails, the pipeline halts  
6970 and an error message appears to alert you to the problem.

6971 – **Use free/open-source software and programming languages.** If

possible, avoid using commercial software, like SPSS or Matlab, and instead use free, open-source software and programming languages, like JASP, Jamovi, R, or Python. This practice will make it easier for others to access, reuse, and verify your work—including yourself!<sup>7</sup>

- **Use version control.** In chapter 13, we introduced the benefits of version control—a great way to save your analysis pipeline incrementally as you build it, allowing you to roll back to a previous version if you accidentally introduce errors.
- **Preserve the computational environment.** Even if your analysis pipeline is entirely reproducible on your own computer, you still need to consider whether it will run on somebody else’s computer, or even your own computer after software updates. You can address this issue by documenting and preserving the computational environment in which the analysis pipeline runs successfully. Various tools are available to help with this, including Docker, Code Ocean, renv (for R), and pip (for Python).<sup>8</sup>

#### 6989 14.3.3 *The reproducibility-collaboration trade-off*

6990 We would love to leave it there and watch you walk off into the sunset  
6991 with a spring in your step and a reproducible report under your arm.

<sup>7</sup> Several of us have libraries of old Matlab code. While discounted licenses are available to students, a full-price software license can be a major barrier to researchers with limited resources. If you move away from Matlab, it’s also terrible to have to ask yourself whether it’s worth the price of another year’s license just to rerun one old analysis.

<sup>8</sup> If you are interested in going in this direction, we recommend Peikert and Brandmaier (2021), which gives an advanced tutorial for complete computational reproducibility using Docker and make as tools to supplement Git and R Markdown.

6992 Unfortunately, we have to admit that writing reproducibly can create a  
6993 few practical difficulties when it comes to collaboration.

6994 A major aspect of collaboration is exchanging comments and inline text  
6995 edits with coauthors. You can do this exchange with R Markdown files  
6996 and Git, but these tools are not as user-friendly as, say, Word or Google  
6997 Docs, and some collaborators will be completely unfamiliar with them.  
6998 Most journals also expect articles to be submitted as Word documents.  
6999 Outputting R Markdown files to Word can often introduce formatting  
7000 issues, especially for moderately complex tables. So, until more user-  
7001 friendly tools are introduced, some compromise between reproducibil-  
7002 ity and collaboration may be necessary. Here are two workflow styles  
7003 for you to consider.

7004 First, the **maximal reproducibility** approach. If your collaborators are  
7005 familiar with R Markdown and you don't mind exchanging comments  
7006 and edits via Git—or if they don't mind giving you lists of comments and  
7007 changes that you implement in the R Markdown document—then you  
7008 can maintain a fully reproducible workflow for your project. The jour-  
7009 nal submission and publication process may still introduce some issues,  
7010 such as incorporating changes made by the copy editor, but at least your  
7011 submitted manuscript (and the preprint you have hopefully posted) will  
7012 be fully reproducible.

7013 Second, the **two worlds** approach. This workflow is a bit clunky, but it  
7014 facilitates collaboration and maintains reproducibility. First, write your  
7015 results section in R Markdown and generate a Word document (see  
7016 appendix C). Then, write the remainder of the manuscript in Word,  
7017 including incorporating comments and changes from collaborators.  
7018 When you have a final version, copy and paste the abstract, introduction,  
7019 methods, and discussion into the R Markdown document.<sup>9</sup> Integrating  
7020 any changes made to the results section back into the R Markdown  
7021 requires a bit more effort, either using manual checking or Word's  
7022 "compare documents" feature.<sup>10</sup> The advantage of this approach is  
7023 that you have a reproducible document and your collaborators have  
7024 not had to deviate from their preferred workflow. Unfortunately, it  
7025 requires more effort from you and is slightly more error-prone than  
7026 the maximal reproducibility approach.

<sup>9</sup> You can also incorporate Google Docs into this workflow—we find that cloud platforms like Docs are especially useful when gathering comments from multiple collaborators on the same document. Unfortunately, you cannot generate a Google Doc from R Markdown, so you will need to import and convert or else copy and paste.

<sup>10</sup> Packages such as `trackdown` (Kothe et al. 2021) could help as well (<https://claudiozandonella.github.io/trackdown>).

#### 7027 *14.4 Writing responsibly*

7028 As a scientific writer, you have both professional and ethical responsi-  
7029 bilities. You must communicate all relevant information about your  
7030 research so as to enable proper evaluation and verification by other sci-  
7031 entists. It is also important not to overstate your findings and calibrate  
7032 your conclusions to the available evidence (Hoekstra and Vazire 2021).

7033 If errors are found in your work, you must respond and correct them  
7034 when possible (Bishop 2018). Finally, you must meet scholarly obliga-  
7035 tions with regards to authorship and citation practices.

#### 7036 *14.4.1 Responsible disclosure and interpretation*

7037 Back in school, we all learned that getting the right answer is not  
7038 enough—you need to demonstrate how you arrived at that answer  
7039 in order to get full marks. The same expectation applies to research  
7040 reports. Don’t just tell the reader what you found, tell them how you  
7041 found it.<sup>11</sup> That means describing the methods in full detail, as well as  
7042 sharing data, materials, and analysis scripts.

7043 In a journal article, you typically have some flexibility in terms of how  
7044 much detail you provide in the main body of the article and how much  
7045 you relegate to the supplementary information. Readers have different  
7046 needs; some may just want to know the highlights, and some will need  
7047 detailed methodological information in order to replicate your study.

7048 As a rule of thumb, try to make sure there is nothing relegated to the sup-  
7049 plementary information that might surprise the reader. You certainty  
7050 should not use the supplementary information to hide important details  
7051 deliberately or use it as a disorganized dumping ground—the principles  
7052 of clear writing still apply!

<sup>11</sup> It can be easy to overlook important details, especially when you reach the end of a project. Looking back at your study preregistration can be a helpful reminder. Reporting guidelines for different research designs can also provide useful checklists (Appelbaum et al. 2018).

7053 Here are a few more guidelines for responsible writing:

- 7054 – **Don't overclaim.** Scientists often feel they are (and unfortunately,  
7055 often are) evaluated based on the *results* of their research, rather  
7056 than the *quality* of their research. Consequently, it can be tempt-  
7057 ing to make bigger and bolder claims than are really justified by  
7058 the evidence. Think carefully about the limitations of your re-  
7059 search and calibrate your conclusions to the evidence, rather than  
7060 what you wish you were able to claim. Ensure that your con-  
7061 clusions are appropriately stated throughout the manuscript, es-  
7062 pecially in the title and abstract.

- 7063 – **Acknowledge limitations on interpretation and generalizability.**

7064 Even if you calibrate your claims appropriately throughout, there  
7065 are likely specific limitations that are worth discussing, either as  
7066 you introduce the design of the study in the introduction or as you  
7067 interpret it in the discussion section. For example, if your exper-  
7068 iment used one particular manipulation to instantiate a construct  
7069 of interest, you might discuss this limitation and how it might be  
7070 addressed by future work. Think carefully about the limitations  
7071 of your study, state them clearly, and consider how they impact  
7072 your conclusions (Clarke et al. 2023).<sup>12</sup> Discussions of limitations  
7073 are a great point to make an explicit statement about the *generaliz-*

<sup>12</sup> Should you just make your claims more modest and avoid writing about your study's limitations? The balance between claims and limitations is tricky. One way to navigate this issue is to ask yourself, "Is it OK to say X in the abstract of my article, if I later go on to say state a limitation relevant to that claim, or will the reader feel tricked?"

ability of your findings (see Simons, Shoda, and Lindsay 2017 for guidance about these kinds of “Constraints on Generality” statements).

– **Discuss, don’t debate.** The purpose of the discussion section is to help the reader interpret your research. Importantly, a journal article is not a debate—don’t feel the need to argue dogmatically for a particular position or interpretation. You should discuss the strengths and weaknesses of the evidence, and the relative merits of different interpretations. For example, perhaps there is a potential confounding variable that you were unable to eliminate with your research design. The reader might be able to spot this themselves, but regardless, it’s your responsibility to highlight it. Perhaps on balance you think the confound is unlikely to explain the results—that’s fine, but you need to explain your reasoning to the reader.

– **Disclose conflicts of interest and funding.** Researchers are usually personally invested in the outcomes of their research, and this investment can lead to bias (for example, overclaiming or selective reporting). But sometimes your potential personal gains from a piece of research rise above a threshold and are considered **conflicts of interest**. Where this threshold lies is not always com-

7095 pletely clear. The most obvious conflicts of interest occur when  
7096 you stand to benefit financially from the outcomes of your re-  
7097 search (for example, a drug developer evaluating their own drug).

7098 If you are in doubt about whether you have a potential conflict of  
7099 interest, then you should disclose it. You should also disclose any  
7100 funding you received for the research, partly because this is often  
7101 a requirement of the funder and partly because it may represent  
7102 a conflict of interest, for example, if the funder has a particular  
7103 stake in the outcome of the research. To avoid ambiguity, you  
7104 should also disclose when you do *not* have a conflict of interest or  
7105 funding to declare.

- 7106 – **Report transparently.** In chapter 11, you learned about the prob-  
7107 lem of selective reporting and how this practice can bias the re-  
7108 search literature. There are several ways to avoid this issue in your  
7109 own work. First, assuming you *have* reported everything, include  
7110 a statement in the methods section that explicitly says so. A state-  
7111 ment suggested by Simmons, Nelson, and Simonsohn (2012) is  
7112 “We report how we determined our sample size, all data exclu-  
7113 sions (if any), all manipulations, and all measures in the study.”  
7114 If you have preregistered your study, clearly link to the preregis-  
7115 tration and state whether you deviated from your original plan.  
7116 You can include a detailed preregistration disclosure table in the

7117 supplementary information and highlight any major deviations in  
7118 the methods section. In the results section, clearly identify (e.g.,  
7119 with subheadings) which analyses were preplanned and included  
7120 in the preregistration (confirmatory) and which were not planned  
7121 (exploratory).

#### 7122 14.4.2 Responsible handling of errors

7123 It is not your responsibility to never make mistakes. But it *is* your re-  
7124 sponsibility to respond to errors in a timely, transparent, and professional  
7125 manner (Bishop 2018).<sup>13</sup> Regardless of how the error was identified  
7126 (e.g., by yourself or by a reader), we recommend contacting the jour-  
7127 nal and requesting that they publish a correction statement (sometimes  
7128 called an **erratum**). Several of us have corrected papers in the past. If  
7129 the error is serious and cannot be fixed, you should consider retracting  
7130 the article.

7131 A correction/retraction statement should include the following infor-  
7132 mation:

- 7133 1. **Acknowledge the error.** Be clear that an error has occurred.
- 7134 2. **Describe the error.** Readers need to know the exact nature of the  
7135 error.

<sup>13</sup> As jazz musician Miles Davis once said, “If you hit a wrong note, it’s the next note that you play that determines if it’s good or bad.”

- 7136 3. **Describe the implications of the error.** Readers need to know how  
7137 the error might affect their interpretation of the results.
- 7138 4. **Describe how the error occurred.** Knowing how the error hap-  
7139 pened may help others avoid the same error.
- 7140 5. **Describe what you have done to address the error.** Others may  
7141 learn from solutions you've implemented.
- 7142 6. **Acknowledge the person who identified the error.** Identifying er-  
7143 rors can take a lot of work; if the person is willing to be identified,  
7144 give credit where credit is due.

### ACCIDENT REPORT

In 2018, at a crucial stage of her career, Julia Strand published an important study in the prestigious journal *Psychonomic Bulletin & Review*. She presented the work at conferences and received additional funding to do follow-up studies. But several months later, her team found that they could not replicate the result.

Puzzled, she began searching for the cause of the discrepant results. Eventually, she found the culprit—a programming error. As she sat staring at her computer in horror, she realized that it was unlikely anyone else would ever find the bug. Hiding the error must have seemed like the easiest thing to do.

But she did the right thing. She spent the next day informing her students, her coauthors, the funding officer, the department chair overseeing

her tenure review, and the journal—to initiate a retraction of the article. And … it didn’t ruin her career. Everybody was understanding and appreciated that she was doing the right thing. The journal corrected the article. She didn’t lose her grant. She got tenure. And a lot of scientists, including us, admire her for what she did.

Honest mistakes happen—it’s how you respond to them that matters (Strand 2021). In fact, survey research with both scientists and the general public suggests that scientists’ reputations are built on the perception that they try to “get it right,” not just to “be right” (Ebersole, Axt, and Nosek 2016).

7146

#### *7147 14.4.3 Responsible citation*

7148 Citing prior work that your study builds upon ensures that researchers  
7149 receive credit for their contributions and helps readers to verify the basis  
7150 of your claims. You should certainly avoid copying the work of others  
7151 and presenting it as your own (see chapter 4 for more on plagiarism).  
7152 Try to be explicit about why you are citing a source. For example, does  
7153 it provide evidence to support your point? Is it a review paper that gives  
7154 the reader useful background? Or is it a description of a theory you are  
7155 testing?

7156 Make sure you read articles before you cite them. Stang, Jonas, and

7157 Poole (2018) reports a cautionary tale in which a commentary criticizing  
7158 a methodological tool was frequently cited as *supporting* the use of  
7159 that tool! It seems that many authors had not read the paper they were  
7160 citing, which is both misleading and embarrassing.

7161 Try to avoid selective or uncritical citation. It is misleading to cite only  
7162 research that supports your argument and ignoring research that doesn't.  
7163 You should provide a balanced account of prior work, including contra-  
7164 dictory evidence. Make sure to evaluate and integrate evidence from  
7165 prior studies, rather than simply describing them. Remember—every  
7166 study has limitations.

#### 7167 14.4.4 Responsible authorship practices

7168 It is an ethical responsibility to credit the individuals who worked on a  
7169 research project—so that they can reap the benefits if the work is influ-  
7170 ential, but also so that they can take responsibility for errors.<sup>14</sup>

7171 Currently in academia, the *authorship model* is dominant. Under this  
7172 model, authorship and authorship order are important signals about re-  
7173 searchers's contributions to a project. It is generally expected that to  
7174 qualify for authorship, an individual should have made a substantial  
7175 contribution to the research (e.g., design, data collection, analysis) and

<sup>14</sup> In 1975, physicist and mathematician Jack H. Hetherington wrote a paper he intended to submit to the journal *Physical Review Letters*. We're not sure why, but Hetherington wrote the paper in first person plural (i.e., referring to himself as "we" rather than "I"). He subsequently discovered that the journal would not accept the use of "we" for single-authored articles. Hetherington had painstakingly tapped out the article on his typewriter, an exercise he was not keen to repeat. Instead, he opted for a less taxing solution and named his cat—a feline by the name of F. D. C. Willard—as a coauthor. The paper was accepted and published (Hetherington and Willard 1975).

7176 assisted with writing the research report, and that they take joint re-  
7177 sponsibility for the research along with the other coauthors. Individuals  
7178 who worked on the project who do not reach this threshold are instead  
7179 mentioned in a separate acknowledgements section and not considered  
7180 authors.

7181 **Authorship order** is often understood to signal the nature and extent  
7182 of an author's contribution. In psychology (and neighboring disci-  
7183 plines), the first author and last author are typically the project leaders.  
7184 Typically—though certainly not always!—the first author is a junior  
7185 colleague who implements the project and the last author is a senior  
7186 colleague who supervises the project.

7187 It has been argued that the authorship model should be replaced with  
7188 a more inclusive *contributorship* model in which all individuals who  
7189 worked on the project are acknowledged as “contributors.” Unlike the  
7190 authorship model, there is no arbitrary threshold for contributorship.  
7191 The actual contributions of each individual are explicitly described,  
7192 rather than relying on the implicit conventions of authorship order.  
7193 The contributorship model may facilitate collaboration and ensure  
7194 student assistants are properly credited.

7195 You will probably find that most journals still expect you to use the  
7196 authorship model. Nevertheless, it is usually possible—and sometimes

7197 required—to include a contributorship statement in your article that  
7198 describes what everybody did. For example, the CREDIT taxonomy  
7199 provides a structured taxonomy of research tasks, making for uniform  
7200 contributorship reporting.<sup>15</sup>

7201 Because authorship is such an important signal in academia, it's impor-  
7202 tant to agree on an authorship plan with your collaborators (particularly  
7203 who will be the first and last authors) as early as possible.<sup>16</sup>

## 7204 14.5 *Chapter summary: Writing*

7205 Writing a scientific article can be a rewarding endpoint for the process  
7206 of doing experimental research. But writing is a craft, and writing  
7207 clearly—especially about complex and technical topics—can require  
7208 substantial practice and many drafts. Further, writing about research  
7209 comes with ethical and professional responsibilities that are different  
7210 than the burdens of other kinds of writing. A scientific author must  
7211 work to ensure the reproducibility of their findings and report on  
7212 those findings responsibly, noting limitations and weaknesses as well as  
7213 strengths.

<sup>15</sup> For larger projects, the tool Tenzing allows for CREDIT statements to be generated automatically from standardized forms (Holcombe et al. 2020).

<sup>16</sup> If you find yourself in a situation where all authors have contributed equally, you may have to draw inspiration from historical examples and determine authorship order based on a 25-game croquet series (Hassell and May 1974); rock, paper, scissors (Kupfer, Webbeking, and Franklin 2004); or a brownie bake-off (Young and Young 1992). Alternatively, you can adopt the method of Lakens, Scheel, and Isager (2018) and randomize the authorship order in R!



## DISCUSSION QUESTIONS

1. Find a writing buddy and exchange feedback on a short piece of writing (the abstract of a paper in progress, a conference abstract, or even a class project proposal would be good examples). Think about how to improve each other's writing using the advice offered in this chapter.
2. Identify a published research article with openly available data and see if you can reproduce an analysis in their paper by recovering the exact numerical values they report. You can find support for this exercise at the Social Science Reproduction Platform (<https://www.socialescience-reproduction.org>) or ReproHack (<https://www.reprohack.org>). Discuss with a friend what challenges you faced in this exercise and how they might be avoided in your own work.

7214



## READINGS

- Zinsser, William (2006). *On Writing Well: The Classic Guide to Writing Nonfiction*. 7th ed. Harper Collins.
- Gernsbacher, Morton Ann (2018). “Writing Empirical Articles: Transparency, Reproducibility, Clarity, and Memorability.” *Advances in Methods and Practices in Psychological Science* 1 (3): 403–414. <https://doi.org/10.1177/2515245918754485>.

7215

## 7216 References

- Appelbaum, Mark, Harris Cooper, Rex B. Kline, Evan Mayo-Wilson, Arthur M. Nezu, and Stephen M. Rao. 2018. “Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report.” *American Psychologist* 73 (1): 3. <https://doi.org/10.1037/amp0000191>.
- Barnett, Adrian, and Zoe Doubleday. 2020. “The Growth of Acronyms in the Scientific Literature.” *eLife* 9 (July):e60080. <https://doi.org/10.7554/eLife.60080>.
- Bishop, D. V. M. 2018. “Fallibility in Science: Responding to Errors in the Work of Oneself and Others.” *Advances in Methods and Practices in Psychological Science* 1 (3): 432–438. <https://doi.org/10.1177/2515245918776632>.
- Clarke, Beth, Lindsay Alley, Sakshi Ghai, Jessica Kay Flake, Julia M. Rohrer, Joseph P. Simmons, Sarah R. Schiavone, and Simine Vazire. 2023. “Looking Our Limitations in the Eye: A Tutorial for Writing about Research Limitations in Psychology.” PsyArXiv. <https://doi.org/10.31234/osf.io/386bh>.
- Cutler, Anne. 1994. “The Perception of Rhythm in Language.” *Cognition* 50:79–81.
- Doumont, Jean-Luc. 2009. *Trees, Maps, and Theorems*. Principiae.
- Ebersole, Charles R., Jordan R. Axt, and Brian A. Nosek. 2016. “Scientists’ Reputations Are Based on Getting It Right, Not Being Right.” *PLoS Biology* 14 (5): e1002460.
- Gernsbacher, Morton Ann. 2018. “Writing Empirical Articles: Transparency, Reproducibility, Clarity, and Memorability.” *Advances in Methods and*

*Practices in Psychological Science* 1 (3): 403–414. <https://doi.org/10.1177/2515245918754485>.

Hardwicke, Tom E., Maya B. Mathur, Kyle Earl MacDonald, Gustav Nilsonne, George Christopher Banks, Mallory Kidwell, Alicia Hofelich Mohr, et al. 2018. “Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal Cognition.” *Royal Society Open Science* 5. <https://doi.org/10.1098/rsos.180448>.

Hassell, M. P., and R. M. May. 1974. “Aggregation of Predators and Insect Parasites and Its Effect on Stability.” *Journal of Animal Ecology* 43 (2): 567–594. <https://doi.org/10.2307/3384>.

Heard, Stephen B. 2016. *The Scientist’s Guide to Writing: How to Write More Easily and Effectively Throughout Your Scientific Career*. Princeton University Press.

Hetherington, J. H., and F. D. C. Willard. 1975. “Two-, Three-, and Four-Atom Exchange Effects in bcc  ${}^3\text{He}$ .” *Physical Review Letters* 35 (21): 1442–1444. <https://doi.org/10.1103/PhysRevLett.35.1442>.

Heycke, Tobias, and Lisa Spitzer. 2019. “Screen Recordings as a Tool to Document Computer Assisted Data Collection Procedures.” *Psychologica Belgica* 59 (1): 269–280. <https://doi.org/10.5334/pb.490>.

Hoekstra, Rink, and Simine Vazire. 2021. “Aspiring to Greater Intellectual Humility in Science.” *Nature Human Behaviour* 5 (12): 1602–1607.

Holcombe, Alex O., Marton Kovacs, Frederik Aust, and Balazs Aczel. 2020. “Documenting Contributions to Scholarly Articles Using CRediT and Tenzing.” Edited by Cassidy R. Sugimoto. *PLOS ONE* 15 (12): e0244611. <https://doi.org/10.1371/journal.pone.0244611>.

- King, Stephen. 2000. *On Writing: A Memoir of the Craft*. Scribner.
- Knuth, Donald Ervin. 1992. *Literate Programming*. no. 27. Center for the Study of Language and Information.
- Kothe, Emily, Claudio Zandonella Callegher, Filippo Gambarota, Janosch Linkersdörfer, and Mathew Ling. 2021. *trackdown: Collaborative Writing and Editing of r Markdown (or Quarto / Sweave) Documents in Google Drive*. <https://doi.org/10.5281/zenodo.5772942>.
- Kupfer, John A., Amy L. Webbeking, and Scott B. Franklin. 2004. “Forest Fragmentation Affects Early Successional Patterns on Shifting Cultivation Fields Near Indian Church, Belize.” *Agriculture, Ecosystems & Environment* 103 (3): 509–518. <https://doi.org/10.1016/j.agee.2003.11.011>.
- 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): 259–269. <https://doi.org/10.1177/2515245918770963>.
- Lewis, Michael. 2016. *The Undoing Project: A Friendship That Changed the World*. Penguin UK.
- Nuijten, Michèle B., Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. 2016. “The Prevalence of Statistical Reporting Errors in Psychology (1985–2013).” *Behavior Research Methods* 48 (4): 1205–1226.
- Peikert, Aaron, and Andreas M. Brandmaier. 2021. “A Reproducible Data Analysis Workflow with r Markdown, Git, Make, and Docker.” *Quantitative and Computational Methods in Behavioral Sciences*, 1–27.
- Savage, Van, and Pamela Yeh. 2019. “Novelist Cormac McCarthy’s Tips on

- How to Write a Great Science Paper.” *Nature* 574 (7777): 441–443.
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2012. “A 21 Word Solution.” *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2160588>.
- Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay. 2017. “Constraints on Generality (COG): A Proposed Addition to All Empirical Papers.” *Perspectives on Psychological Science* 12 (6): 1123–1128.
- Stang, Andreas, Stephan Jonas, and Charles Poole. 2018. “Case Study in Major Quotation Errors: A Critical Commentary on the Newcastle–Ottawa Scale.” *European Journal of Epidemiology* 33 (11): 1025–1031. <https://doi.org/10.1007/s10654-018-0443-3>.
- Strand, Julia. 2021. “Error Tight: Exercises for Lab Groups to Prevent Research Mistakes.” PsyArXiv. <https://doi.org/10.31234/osf.io/rsn5y>.
- Young, Helen J., and Truman P. Young. 1992. “Alternative Outcomes of Natural and Experimental High Pollen Loads.” *Ecology* 73 (2): 639–647. <https://doi.org/10.2307/1940770>.
- Zinsser, William. 2006. *On Writing Well: The Classic Guide to Writing Nonfiction*. 7th ed. HarperCollins.

## <sup>7222</sup> 15 VISUALIZATION

### LEARNING GOALS

- Analyze the principles behind informative visualizations
- Incorporate visualization into an analysis workflow
- Learn to make “the design plot”: a standard visualization of experimental data
- Select different visualizations of variability and distribution
- Connect visualization concepts to measurement principles

<sup>7223</sup>

<sup>7224</sup> What makes visualizations so useful, and what role do they play in  
<sup>7225</sup> the experimenter’s toolkit? Simply put, data visualization is the act of  
<sup>7226</sup> “making the invisible visible.” Our visual systems are remarkably pow-  
<sup>7227</sup> erful pattern detectors, and relationships that aren’t at all clear when  
<sup>7228</sup> scanning through rows of raw data can immediately jump out at us when  
<sup>7229</sup> presented in an appropriate graphical form (Zacks and Franconeri 2020).  
<sup>7230</sup> Good visualizations aim to deliberately harness this power and put it to  
<sup>7231</sup> work at every stage of the research process, from the quick sanity checks

<sub>7232</sub> we run when first reading in our data to the publication-quality figures

<sub>7233</sub> we design when we are ready to communicate our findings.

<sub>7234</sub> Yet, our powerful pattern detectors can also be a liability; if we're not

<sub>7235</sub> careful, we can easily be fooled into seeing patterns that are unreliable

<sub>7236</sub> or even misleading. As psychology moves into an era of bigger data and

<sub>7237</sub> more complex behaviors, we become increasingly reliant on **data visu-**

<sub>7238</sub> **alization literacy** (Börner, Bueckle, and Ginda 2019) to make sense of

<sub>7239</sub> what is going on. Further, as a researcher reporting about your data, cre-

<sub>7240</sub> ating appropriate visualizations that are aligned with your analyses (as

<sub>7241</sub> well as your design and preregistration) is an important part of TRANS-

<sub>7242</sub> PARENCEY and BIAS REDUCTION in your reporting.



### *Mapping a pandemic*

In 1854, a deadly outbreak of cholera was sweeping through London.

The scientific consensus at the time was that diseases like cholera spread through breathing poisonous and foul-smelling vapors, an idea known as the “miasma theory” (Halliday 2001). An obstetrician and anesthesiologist named John Snow, however, had proposed an alternative theory: rather than spreading through foul air, he thought that cholera was spreading through a polluted water supply (Snow 1855). To make a public case for this idea, he started counting cholera deaths. He marked each

case on a map of the area and indicated the locations of the water pumps for reference. Furthermore, a line could be drawn representing the region that was closest to each water pump, a technique that is now known as a Voronoi diagram ([https://en.wikipedia.org/wiki/Voronoi\\_diagram](https://en.wikipedia.org/wiki/Voronoi_diagram)). The resulting illustration clearly reveals that cases clustered around an area called Golden Square, which received water from a pump on Broad Street (figure 15.1). Although the precise causal role of these maps in Snow's own thinking is disputed, and it is likely that he produced them well after the incident (Brody et al. 2000), they nonetheless played a significant role in the history of data visualization (Friendly and Wainer 2021).

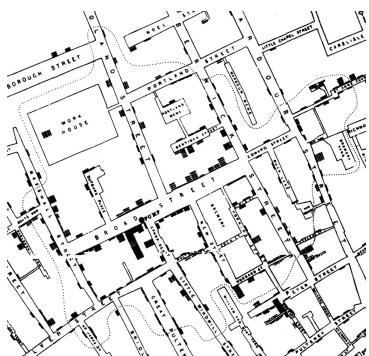


Figure 15.1  
Mapping out a cholera epidemic (Snow 1854). The dotted line shows the region for which Broad Street pump is nearest.

Nearly two centuries later, as the COVID-19 pandemic swept through the world, governmental agencies like the CDC produced maps of the outbreak that became much more familiar (figure 15.2).

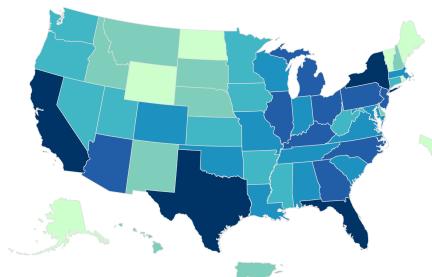


Figure 15.2

A map showing the counts of COVID hospitalizations by state since August 2020 as of January 2024 (from the CDC COVID Data Tracker, <https://covid.cdc.gov/covid-data-tracker>). Usage does not constitute endorsement or recommendation by the US Government, Department of Health and Human Services, or Centers for Disease Control and Prevention.

These maps make abstract statistics visible: By assigning higher cumulative case rates to darker colors, we can see at a glance which areas have been most affected. And we're not limited by the spatial layout of a map. We're now also used to seeing the horizontal axis correspond to *time* and the vertical axis correspond to some value at that time. Curves like the following, showing the weekly counts of new cases, allow us to see other patterns, like the *rate of change*. Even though more and more cases accumulate every day, we can see at a glance the different “waves” of cases and when they peaked (figure 15.3).

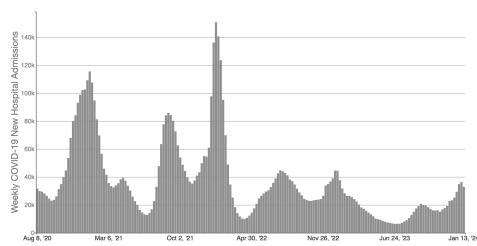


Figure 15.3

Weekly counts of new reported COVID hospital admissions in the US between August 2020 and January 2024 (from the CDC COVID Data Tracker, <https://covid.cdc.gov/covid-data-tracker>). Usage does not constitute endorsement or recommendation by the US Government, Department of Health and Human Services, or Centers for Disease Control and Prevention.

While these visualizations capture purely descriptive statistics, we often want our visualizations to answer more specific questions. For example, we may ask about the effectiveness of vaccinations: How do case rates differ across vaccinated and unvaccinated populations? In this case, we may talk about “breaking out” a curve by some other variable, like vaccination status (figure 15.4).

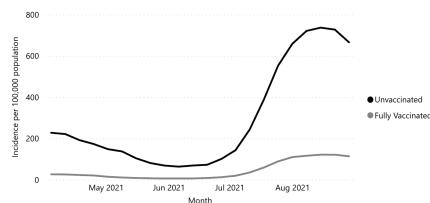


Figure 15.4

Rates of COVID cases by vaccination status (from the CDC COVID Data Tracker, <https://covid.cdc.gov/covid-data-tracker>). Usage does not constitute endorsement or recommendation by the US Government, Department of Health and Human Services, or Centers for Disease Control and Prevention.

From this visualization, we can see that unvaccinated individuals are about six times more likely to test positive. At the same time, these visualizations were produced using *observational* data, which makes it challenging to draw causal inferences. For example, people were not randomly assigned to vaccination conditions, and those who have avoided vaccinations may differ in other ways than those who sought out vaccinations. Additionally, you may have noticed that these visualizations typically do not give a sense of the raw data, the sample sizes of each group, or uncertainty about the estimates. In this chapter, we will explore how to use visualizations to communicate the results of carefully controlled psychology experiments, which license stronger causal inferences.

## 15.1 Basic principles of (confirmatory) visualization

In this section, we begin by introducing a few simple guidelines to keep in mind when making informative visualizations in the context of experimental psychology.<sup>1</sup> Remember that our needs may be distinct from other fields, such as journalism or public policy. You may have seen beautiful and engaging full-page graphics with small print and a wealth of information. The art of designing and producing these graphics is typically known as **infoviz** and should be distinguished from what we call **statistical visualization** (Gelman and Unwin 2013).

Roughly, infoviz aims to construct rich and immersive worlds to visually explore: a reader can spend hours pouring over the most intricate graphics and continue to find new and intriguing patterns. Statistical visualization, on the other hand, aims to crisply convey the logic of a specific inference at a glance. These visualizations are the production-ready figures that anchor the results section of a paper and accompany the key, preregistered analyses of interest. In this section, we review several basic principles of making statistical visualizations. We then return below to the role of visualization in more exploratory analyses.

<sup>1</sup> For the purposes of understanding the examples in this chapter, it should be sufficient to work through the tutorials on data manipulation and visualization in appendix D and appendix E.

7265 15.1.1 Principle 1: Show the design

7266 There are so many different kinds of graphs (bar graphs, line graphs,  
7267 scatter plots, and pie charts) and so many different possible attributes of  
7268 those graphs (colors, sizes, line types). How do we begin to decide how  
7269 to navigate these decisions? The first principle guiding good statistical  
7270 visualizations is to *show the design* of your experiment.

7271 The first confirmatory plot you should have in mind for your exper-  
7272 iment is the **design plot**. Analogous to the “default” or “saturated”  
7273 model in chapter 7, the design plot should show the key dependent  
7274 variable of the experiment, broken down by all of the key manipula-  
7275 tions. Critically, design plots should neither omit particular manipu-  
7276 lations because they didn’t yield an effect nor include extra covariates  
7277 because they seemed interesting after looking at the data. Both of these  
7278 steps are the visual analogue of *p*-hacking! Instead, the design plot is the  
7279 “preregistered analysis” of your visualization: it illustrates a first look at  
7280 the estimated causal effects from your experimental manipulations. In  
7281 the words of Coppock (2019), “Visualize as You Randomize”!

7282 It can sometimes be a challenge to represent the full pattern of manipula-  
7283 tions from an experiment in a single plot. Below we give some tricks for  
7284 maximizing the legible information in your plot. But if you have tried  
7285 these and your design plot still looks crowded and messy, that could be

7286 an indication that your experiment is manipulating too many things at

7287 once!

7288 There are strong (unwritten) conventions about how your confirmatory

7289 analysis is expected to map onto graphical elements, and following these

7290 conventions can minimize confusion. Start with the variables you ma-

7291 nipulate, and make sure they are clearly visible. Conventionally, the

7292 primary manipulation of interest (e.g., condition) goes on the x-axis,

7293 and the primary measurement of interest (e.g., responses) goes on the y-

7294 axis. Other critical variables of interest (e.g., secondary manipulations

7295 and demographics) are then assigned to “visual variables” (e.g., color,

7296 shape, or size).

### CODE

The visualization library `ggplot` (see appendix E) makes the mapping of variables in the data to visual data. Part of a `ggplot` call is an `aes` (short for aesthetics) mapping:

```
aes(  
  x = ...,  
  y = ...,  
  color = ...,  
  linetype = ...,  
)
```

The aesthetics argument serves as a statement of how data are mapped to “marks” on the plot. This transparent mapping makes it very easy to explore different plot types by changing that `aes()` statement, as we’ll see below.

7298

7299 As an example, we will consider the data from Stiller, Goodman, and  
7300 Frank (2015) that we explored back in chapter 7. Because this experi-  
7301 ment was a developmental study, the primary independent variable of  
7302 interest was the age group of participants (ages two, three, or four). So  
7303 age gets assigned to the horizontal (x) axis. The dependent variable  
7304 is accuracy: the proportion of trials that a participant made the cor-  
7305 rect response (out of four trials). So accuracy goes on the vertical (y)  
7306 axis. Now, we have two other variables that we might want to show:  
7307 the condition (experimental vs control) and the type of stimuli (houses,  
7308 beds, and plates of pasta). When we think about it, though, only condi-  
7309 tion is central to exposing the design. While we might be interested  
7310 in whether some types of stimuli are systematically easier or harder  
7311 than others, condition is more central for understanding the *logic* of the  
7312 study.

## CODE

As a reminder, here's our code for loading the Stiller, Goodman, and Frank (2015) data:

```
repo <- "https://raw.githubusercontent.com/langcog/experimentology/main"

sgf <- read_csv(file.path(repo, "data/tidyverse/stiller_scales_data.csv")) |>

  mutate(age_group = cut(age, 2:5, include.lowest = TRUE),
         condition = condition |>

    fct_recode("Experimental" = "Label", "Control" = "No Label"))

sgf_cond_means <- sgf |>

  group_by(condition, age_group) |>

  summarize(rating = mean(correct))
```

7313

### 15.1.2 Principle 2: Facilitate comparison

7314 Now that you've mapped elements of your design to the figure's axes,  
 7315 how do you decide which graphical elements to display? You might  
 7316 think: well, in principle, these assignments are all arbitrary anyway. As  
 7317 long as we clearly label our choices, it shouldn't matter whether we  
 7318 use lines, points, bars, colors, textures, or shapes. It's true that there  
 7319 are many ways to show the same data. But being thoughtful about our  
 7320 choices can make it much easier for readers to interpret our findings.  
 7321 The second principle of statistical visualizations is to *facilitate comparison*

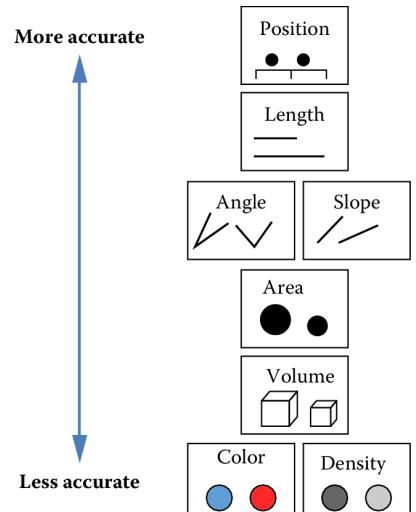


Figure 15.5  
 Principles of visual perception can help guide visualization choices. Based on Mackinlay (1986; see also Cleveland and McGill 1984).

7323 along the dimensions relevant to our scientific questions. It is easier for  
7324 our visual system to accurately compare the location of elements (e.g.,  
7325 noticing that one point is a certain distance away from another) than to  
7326 compare their areas or colors (e.g., noticing that one point is bigger or  
7327 brighter than another). Figure 15.5 shows an ordering of visual variables  
7328 based on how accurate our visual system is in making comparisons.

7329 For example, we *could* start by plotting the accuracy of each age group  
7330 as colors (figure 15.6).

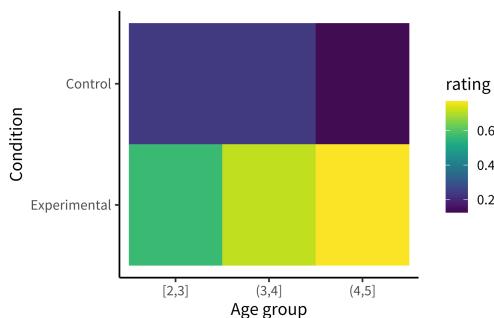


Figure 15.6  
A first visualization of the Stiller, Goodman, and Frank (2015) data.

 CODE

To make this (bad) visualization, we used a `ggplot` function called `geom_tile()`.

```
ggplot(sgf_cond_means, aes(x = age_group, y = condition, fill = rating)) +  
  geom_tile() +  
  labs(x = "Age group", y = "Condition")
```

`geom_tile()` is commonly used to make heat maps ([https://en.wikipedia.org/wiki/Heat\\_map](https://en.wikipedia.org/wiki/Heat_map)): for each value of some pair of variables (x, y), it shows a color representing the magnitude of a third variable (z).

While a heat map is a silly way to visualize the Stiller, Goodman, and Frank (2015) data, consider using `geom_tile()` when you have a pair of continuous variables, each taking a large range of values. In these cases, bar plots and line plots tend to get extremely cluttered, making it hard to see the relationship between the variables. Heat maps help these relationships to pop out as clear “hot” and “cold” regions. For example, in Barnett, Griffiths, and Hawkins (2022), a heatmap was used to show a specific range of parameters where an effect of interest emerged (see figure 15.7).

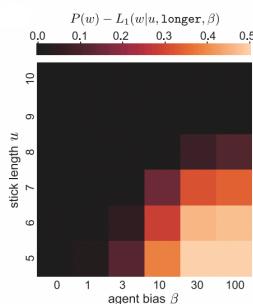


Figure 15.7

A heatmap showing a specific range of continuous parameters where an effect emerged. Barnett, Griffiths, and Hawkins (2022), Figure 3 (licensed under CC BY 4.0).

7332

7333 Or we could plot the accuracy of each age group as sizes/areas (fig-

7334 ure 15.8).

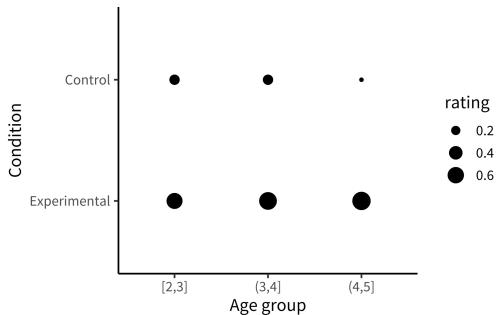


Figure 15.8  
Iterating on the Stiller data using size.

## CODE

To make this (bad) visualization, we mapped the rating DV to the `size` element in our `aes()` call.

```
ggplot(sgf_cond_means, aes(x = age_group, y = condition, size = rating)) +
  geom_point() +
  labs(x = "Age group", y = "Condition")
```

7335

7336 These plots allow us to see that one condition is (qualitatively) bigger  
 7337 than others, but it's hard to tell how much bigger. Additionally, this  
 7338 way of plotting the data places equal emphasis on age and condition, but  
 7339 we may instead have in mind particular contrasts, like the *change* across  
 7340 ages and how that change differs across conditions. An alternative is to  
 7341 show six bars: three on the left showing the “experimental” phase and  
 7342 three on the right showing the “control” phase. Maybe the age groups  
 7343 then are represented as different colors, as in figure 15.9.

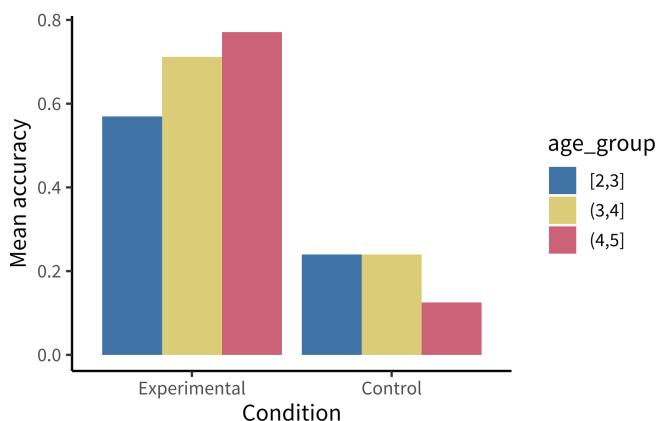


Figure 15.9  
 A bar graph of the Stiller data.

 CODE

We make bar plots using the `ggplot` function `geom_col()`. By default, it creates “stacked” bar plots, where all values associated with the same x value (here, `condition`) get stacked up on top of one another. Stacked bar plots can be useful if, for example, you’re plotting proportions that sum up to 1, or want to show how some big count is broken down into subcategories. It’s also common to use `geom_area()` for this purpose, which connects adjacent regions. But the more common bar plot used in psychology puts the bars next to one another, or “dodges” them. To accomplish this, we use the `position = "dodge"` argument:

```
ggplot(sgf_cond_means, aes(x = condition, y = rating, fill = age_group)) +  
  geom_col(position = "dodge") +  
  labs(x = "Condition", y = "Mean accuracy")
```

7344

7345 This plot is slightly better: it’s easier to compare the heights of bars than  
7346 the “blueness” of squares, and mapping age to color draws our eye to  
7347 those contrasts. However, we can do even better by noticing that our  
7348 experiment was designed to test an *interaction*. That statistic of interest  
7349 is a difference of differences. To what extent does the developmental  
7350 change in performance on the experimental condition differ from de-  
7351 velopmental change in performance on the control condition? Some  
7352 researchers have gotten proficient at reading off interactions from bar  
7353 plots, but they also require a complex set of eye movements. We have

7354 to look at the pattern across the bars on the left, and then jump over  
 7355 to the bars on the right, and implicitly judge one difference against the  
 7356 other: the actual statistic isn't explicitly shown anywhere! What could  
 7357 help facilitate this comparison? Consider the line plot in figure 15.10.

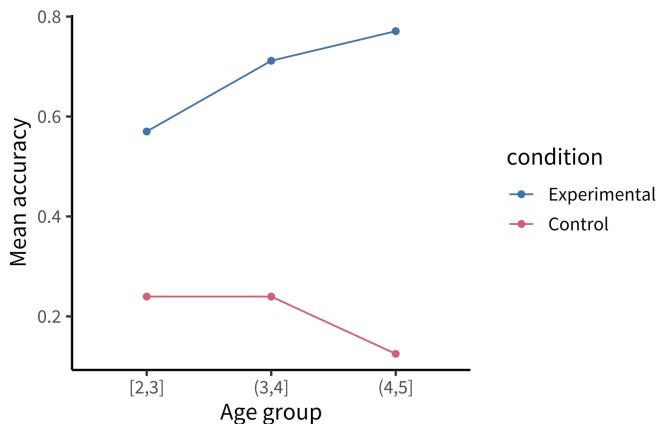


Figure 15.10  
 A line graph of the Stiller data promotes comparison.

## UserCode

Using a combination of `geom_point()` and `geom_line()`:

```
ggplot(sgf_cond_means, aes(x = age_group, y = rating, color = condition)) +  

  geom_point() +  

  geom_line(aes(group = condition)) +  

  labs(x = "Age group", y = "Mean accuracy")
```

7358

7359 The interaction contrast we want to interpret is highlighted visually in  
 7360 this plot. It is much easier to compare slopes of lines than mentally com-  
 7361 pute a difference of differences between bars. Here are a few corollaries  
 7362 of this principle (adapted from a presentation by Karl Broman<sup>2</sup>).

<sup>2</sup> [https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs\\_combined.pdf](https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf)

- 7363 – It is easier to compare values that are *adjacent* to one another. This  
7364 is especially important when there are many different conditions  
7365 included on the same plot. If particular sets of conditions are of  
7366 theoretical interest, place them close to one another. Otherwise,  
7367 sort conditions by a meaningful value (rather than alphabetically,  
7368 which is usually the default for plotting software).
- 7369 – When possible, color-code labels and place them directly next to  
7370 data rather than in a separate legend. Legends force readers to  
7371 glance back and forth to remember what different colors or lines  
7372 mean.
- 7373 – When making histograms or density plots, it is challenging to  
7374 compare distributions when they are placed side by side. Instead,  
7375 facilitate comparison of distributions by vertically aligning them,  
7376 or making them transparent and placed on the same axes.
- 7377 – If the scale makes it hard to see important differences, consider  
7378 transforming the data (e.g., taking the logarithm).
- 7379 – When making bar plots, be very careful about the vertical y-axis.  
7380 A classic “misleading visualization” mistake is to cut off the bot-  
7381 tom of the bars by placing the endpoint of the y-axis at some

arbitrary value near the smallest data point. This is misleading because people interpret bar plots in terms of the relative *area* of the bars (i.e., the amount of ink taken up by the bar), not just their absolute y-values.

- If a key variable from your design is mapped to color, choose the color scale carefully. For example, if the variable is binary or categorical, choose visually distinct colors to maximize contrast (e.g., black, blue, and orange). If the variable is ordinal or continuous, use a color gradient. If there is a natural midpoint (e.g., if some values are negative and some are positive), consider using a diverging scale (e.g., different colors at each extreme). Remember also that a portion of your audience may be colorblind.<sup>3</sup>

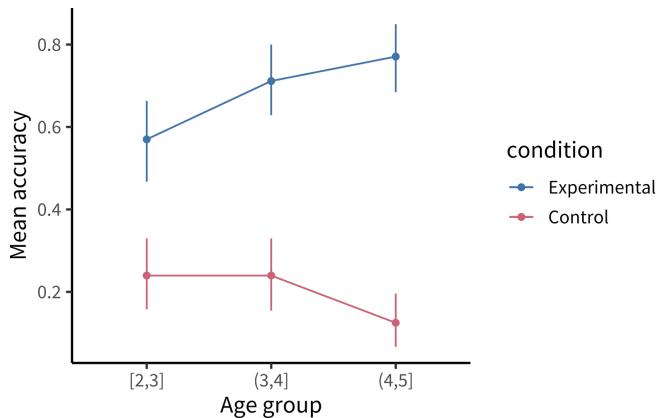
### 15.1.3 Principle 3: Show the data

Looking at older papers, you may be alarmed to notice how little information is contained in the graphs. The worst offenders might show just two bars, representing average values for two conditions. This kind of plot adds very little beyond a sentence in the text reporting the means, but it can also be seriously misleading. It hides real variation in the data, making a noisy effect based on a few data points look the same as a more

<sup>3</sup> Palettes like *viridis* have been designed to be colorblind-friendly and also perceptually uniform (i.e., the perceived difference between 0.1 and 0.2 is approximately the same as the difference between 0.8 and 0.9).

7401 systematic one based on a larger sample. Additionally, it collapses the *dis-*  
 7402 *tribution* of the data, making a multimodal distribution look the same as  
 7403 a unimodal one. The third principle of modern statistical visualization  
 7404 is to *show the data* and visualize variability in some form.

7405 The most minimal form of this principle is to *always include error bars*.<sup>4</sup>  
 7406 Error bars turn a purely descriptive visualization into an inferential one.  
 7407 They represent a minimal form of uncertainty about the possible statis-  
 7408 tics that might have been observed, not just the one that was actually  
 7409 observed. figure 15.11 shows the data with (bootstrapped) error bars.



<sup>4</sup> And be sure to tell the reader what the error bars represent—a 95% confidence interval? A standard error of the mean?—without this information, error bars are hard to interpret (see the DEPTH box below).

Figure 15.11  
 Error bars (95% CIs) added to the Stiller data line graph.

## </> CODE

A common problem arises when we want to add error bars to a dodged bar plot. Naively, we'd expect we could just dodge the error bars in the same way we dodged the bars themselves:

```
geom_col(position = "dodge") +  
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), position = "dodge")
```

But this doesn't work! The rationale is kind of technical, but the width of the error bars is much narrower than the width of the bars, so we need to manually specify how much to dodge the error bars with the `position_dodge()` function:

```
geom_col(position = position_dodge()) +  
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper),  
    position = position_dodge(width = 0.9))
```

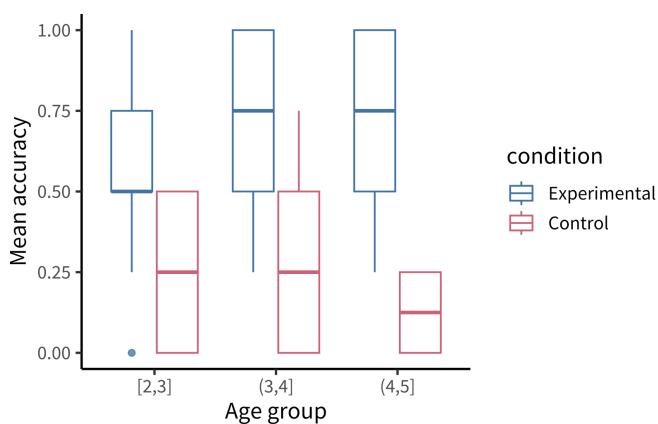
This does the trick!

7411

7412 But we can do even better. By overlaying the distribution of the actual  
7413 data points on the same plot, we can give the reader information about  
7414 not just the statistical inferences but also the underlying data supporting  
7415 those inferences. In the case of the Stiller, Goodman, and Frank (2015)  
7416 study, data points for individual trials are binary (correct or incorrect).  
7417 It's technically possible to show individual responses as dots at 0s and 1s,  
7418 but this doesn't tell us much (we'll just get a big clump of 0s and a big  
7419 clump of 1s). The question to ask yourself when “showing the data”  
7420 is: What are the theoretically meaningful *units* of variation in the data?  
7421 This question is closely related to our discussion of mixed-effects models

in chapter 7, when we considered which random effects we should include. Here, a reader is likely to wonder how much variance was found across *different children* in a given age group. To show such variation, we aggregate to calculate an accuracy score for each participant.<sup>5</sup>

There are many ways of showing the resulting distribution of participant-level data. For example, a boxplot shows the median (a horizontal line) in the center of a box extending from the lower quartile (25%) to the upper quartile (75%). Lines then extend out to the biggest and smallest values (excluding outliers, which are shown as dots). Figure 15.12 gives the boxplots for the Stiller data, which don't look that informative—perhaps because of the coarseness of individual participant averages due to the small number of trials.



<sup>5</sup> While participant-level variation is a good default, the relevant level of aggregation may differ across designs. For example, collective behavior studies may choose to show the data point for each *group*. This choice of unit is also important when generating error bars: if you have a small number of participants but many observations per participant, you are faced with a choice. You may either bootstrap over the flat list of all individual observations (yielding very small error bars), or you may first aggregate within participants (yielding larger error bars that account for the fact that repeated observations from the same participant are not independent).

Figure 15.12  
A boxplot of the Stiller data.

 CODE

In ggplot, we can make box plots using `geom_boxplot()`:

```
geom_boxplot(alpha = 0.8)
```

A common problem to run into is that `geom_boxplot()` requires the variable assigned to `x` to be discrete. If you have discrete levels of a numeric variable (e.g., age groups), make sure you've actually converted that variable to a `factor`. Otherwise, if it's still coded as numeric, `ggplot` will collapse all of the levels together!

7435 It is also common to show the raw data as jittered values with low trans-  
 7436 parency. In figure 15.13, we jitter the points because many participants  
 7437 have the same numbers (e.g., 50%), and if they overlap it is hard to see  
 7438 how many points there are.

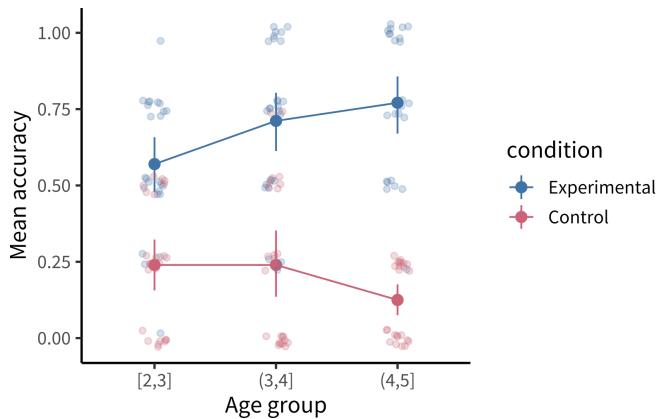


Figure 15.13  
 Jittered points representing the data distribution of the Stiller data.

## CODE

Adding the jittered points is simple using `geom_jitter()`, but we are starting to have a fairly complex plot, so maybe it's worth taking stock of how we get there.

To plot both *condition* means and *participant* means, we need to create two different data frames. Here `sgf_subj_means` is a data frame of means for each participant; `sgf_subj_ci` is a data frame with means and confidence intervals *across* participants. For this purpose, we use the `tidyboot` package (Braginsky and Yurovsky 2018) and the `tidyboot_mean()` function, which gives us bootstrapped 95% confidence intervals for the means.

```
sgf_subj_means <- sgf |>  
  group_by(condition, age_group, subid) |>  
  summarize(rating = mean(correct))  
  
sgf_subj_ci <- sgf_subj_means |>  
  group_by(condition, age_group) |>  
  tidyboot::tidyboot_mean(rating) |>  
  rename(rating = empirical_stat)  
  
ggplot(sgf_subj_ci, aes(x = age_group, y = rating, color = condition)) +  
  geom_pointrange(aes(ymin = ci_lower, ymax = ci_upper)) +  
  geom_line(aes(group = condition)) +  
  geom_jitter(data = sgf_subj_means, alpha = 0.25, width = 0.1, height = .03) +  
  labs(x = "Age group", y = "Mean accuracy")
```

The most noteworthy aspect of this code is that the `geom_jitter()` function doesn't just take a different aesthetic; it also takes a different dataframe altogether! Mixing dataframes can be an important tool for creating complex plots.

7440

7441 Perhaps the format that takes this principle the furthest is the so-called  
7442 raincloud plot (Allen et al. 2019) shown in figure 15.14. A raincloud  
7443 plot combines the raw data (“rain”) with a smoothed density (“cloud”)  
7444 and a boxplot giving the median and quartiles of the distribution.

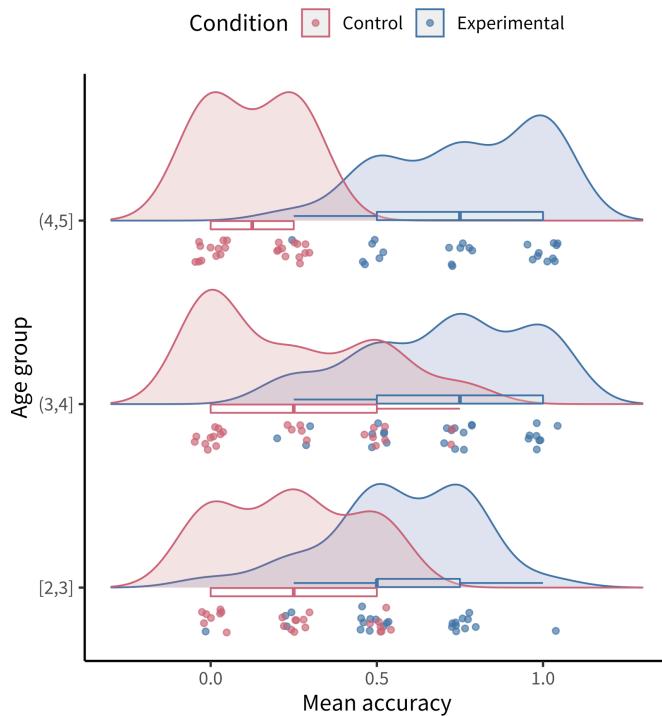


Figure 15.14  
A raincloud plot of the Stiller data.

### CODE

This raincloud plot requires two additional plotting packages: `ggridges` (Wilke 2023) for the densities and `ggstance` (Henry, Wickham, and Chang 2022) for the horizontal boxplots.

```
library(gggridges)
library(ggstance)

ggplot(sgf_subj_means, aes(y = age_group, x = rating, color = condition)) +
  geom_density_ridges(aes(fill = condition), alpha = 0.2, scale = 0.7,
    jittered_points = TRUE, point_alpha = 0.7,
    position = position_raincloud(width = 0.05, height = 0.15,
      ygap = 0.1)) +
  geom_boxplot(width = 0.1, alpha = 0.2, outlier.shape = NA, show.legend = FALSE) +
  scale_y_discrete(expand = expansion(mult = c(0.2, 0.4))) +
  guides(fill = "none", color = guide_legend(reverse = TRUE)) +
  labs(x = "Mean accuracy", y = "Age group", color = "Condition") +
  theme(legend.position = "top")
```

7446

## DEPTH

### *Visualizing uncertainty with error bars*

One common misconception is that error bars are a measure of variance *in the data*, like the standard deviation of the response variable. Instead, they typically represent a measure of precision extracted from the statistical model. In older papers, for example, it was common to use the standard error of the mean (SEM; see chapter 6). Remember that this is not the standard deviation of the data distribution but of the *sampling distribution* of the mean that is being estimated. Given the central limit theorem, which tells us that this sampling distribution is asymptotically normal, it

7447

was straightforward to estimate the standard error analytically using the empirical standard deviation of the data divided by the square root of the sample size: `sd(x) / sqrt(length(x))`. Error bars based on the SEM often looked misleadingly small, as they only represent a 68% interval of the sampling distribution and go to zero quickly as a function of sample size. As a result, it became more common to show the 95% confidence interval instead:  $[-1.96 \times \text{SEM}, 1.96 \times \text{SEM}]$ .

While these analytic equations remain common, an increasingly popular alternative is to *bootstrap* confidence intervals (see the DEPTH box in chapter 6 for more on bootstrapping). The bootstrap is a powerfully generic technique, especially when you want to show error bars for summary statistics that are more complex than means, where we do not have such convenient asymptotic guarantees and “closed-form” equations. An example would be if you’re working with a skewed response variable or a dataset with clear outliers and you want to estimate medians and quartiles.

Or, suppose you want to estimate proportions from categorical data, or a more ad hoc statistic like the AUC (area underneath the curve) in a hierarchical design where it is not clear how to aggregate across items or participants in a mixed-effects model. Analytic estimators of confidence intervals can in principle be derived for these statistics, subject to different assumptions, but it is often more transparent and reliable in practice to use the bootstrap. As long as you can write a code snippet to compute a value from a dataset, you can use the bootstrap.

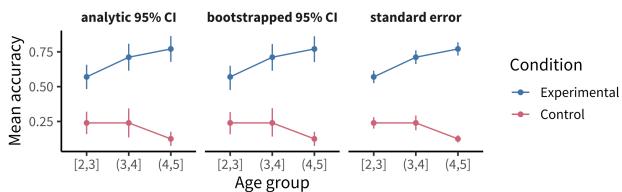


Figure 15.15

Three different error bars for the Stiller data: bootstrapped 95% confidence intervals (left), standard error of the mean (middle), and analytically computed confidence intervals (right).

As we can see, the bootstrapped 95% CI looks similar to the analytic 95% CI derived from the standard error, except the upper and lower limits are slightly asymmetric (reflecting outliers in one direction or another). Of course, the bootstrap is not a silver bullet and can be abused in particularly small samples. This is because the bootstrap is fundamentally limited to the sample we run it on. It can be expected to be reasonably accurate if the sample is reasonably representative of the population. But at the end of the day, as they say, “There’s no such thing as a free lunch.” In other words, we cannot magically pull more information out of a small sample without making additional assumptions about the data generating process.

7449

#### 7450 15.1.1 Principle 4: Maximize information, minimize ink

7451 Now that we have the basic graphical elements in place to show our  
 7452 design and data, it might seem like the rest is purely a matter of aesthetic  
 7453 preference, like choosing a pretty color scheme or font. Not so.

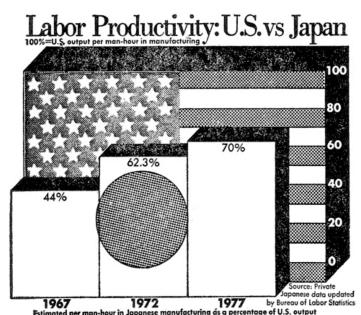


Figure 15.16

This figure uses a lot of ink to show three numbers, for a “ddi” of 0.2 (from the *Washington Post*, 1978; see Wainer (1984) for other examples).

7454 There are well-founded principles to make the difference between an  
 7455 effective visualization and a confusing or obfuscating one. Simply put,  
 7456 we should try to use the simplest possible presentation of the maxi-  
 7457 mal amount of information: we should maximize the “data-ink ratio.”  
 7458 To calculate the amount of information shown, Tufte (2001) suggests  
 7459 a measure called the “data density index,” the “numbers plotted per  
 7460 square inch.” The worst offenders have a very low density while also  
 7461 using a lot of excess ink (e.g., figure 15.16 and figure 15.17)

7462 The defaults in modern visualization libraries like ggplot prevent  
 7463 some of the worst offenses but are still often suboptimal. For example:  
 7464 consider whether the visual complexity introduced by the default grey  
 7465 background and grid lines in figure 15.18 is justified, or whether a more  
 7466 minimal theme would be sufficient.<sup>6</sup>

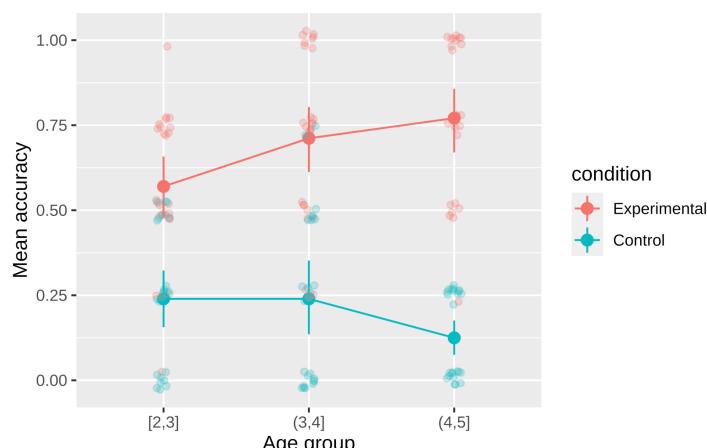


Figure 15.18  
 Standard “gray”-themed Stiller figure.

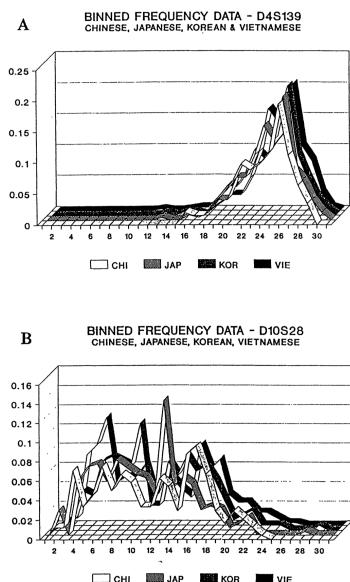


Figure 15.17  
 This figure uses complicated 3D ribbons to compare distributions across four countries (from Roeder 1994). How could the same data have been presented more legibly?

<sup>6</sup> See the `ggthemes` package (Arnold 2023) for a good collection of themes.

7467 Figure 15.19 shows a slightly more “styled” version of the same plot

7468 with labels directly on the plot and a lighter-weight theme.

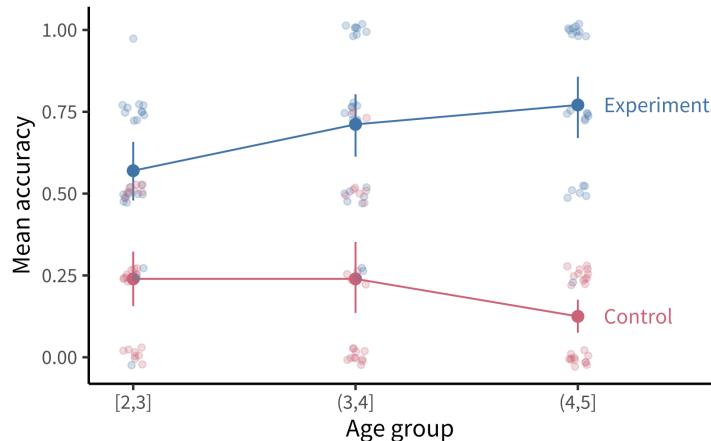


Figure 15.19  
Custom-themed Stiller figure with direct labels.

## CODE

To produce the plot in figure 15.19, we’ve added a few styling elements including:

- The nice and minimal custom theme, with a larger font size.
- A more accessible color palette (`scale_color_ptol()`) from the `ggthemes` package (Arnold 2023).
- Direct labels using `geom_dl()` from the `directlabels` package (Hocking 2023).

```
geom_dl(aes(label = condition), method = list("last.points", dl.trans(x = x + 0.5)))
```

7469

7470 Here are a few final tips for making good confirmatory visualizations:

- 7471 – Make sure the font size of all text in your figures is legible and no smaller than other text in your paper (e.g., 10 pt). This change

7472

7473 may require, for example, making the axis breaks sparser, rotating  
7474 text, or changing the aspect ratio of the figure.

- 7475 – Another important tool to keep in your visualization arsenal is the  
7476 **facet plot**. When your experimental design becomes more com-  
7477 plex, consider breaking variables out into a *grid* of facets instead  
7478 of packing more and more colors and line-styles onto the same  
7479 axis. In other words, while higher information density is typi-  
7480 cally a good thing, you want to aim for the sweet spot before it  
7481 becomes too dense and confusing. Remember principle 2. When  
7482 there is too much going on in every square inch, it is difficult to  
7483 guide your reader’s eye to the comparisons that actually matter,  
7484 and spreading it out across facets gives you additional control over  
7485 the salient patterns.

- 7486 – Sometimes these principles come into conflict, and you may need  
7487 to prioritize legibility over, for example, showing all of the data.  
7488 For example, suppose there is an outlier orders of magnitude away  
7489 from the summary statistics. If the axis limits are zoomed out to  
7490 show that point, then most of the plot will be blank space! It is  
7491 reasonable to decide that it is not worth compressing the key sta-  
7492 tistical question of your visualization into the bottom centimeter  
7493 just to show one point. It may suffice to truncate the axes and

7494 note in the caption that a single point was excluded.

7495 – Fix the axis labels! A common mistake is to keep the default  
7496 shorthand you used to name variables in your plotting software  
7497 instead of more descriptive labels (e.g., “RT” instead of “Reaction  
7498 Time”). Use consistent terminology for different manipulations  
7499 and measures in the main text and figures. If anything might be  
7500 unclear in the figure, explain it in the caption.

7501 – Different audiences may require different levels of detail. Some-  
7502 times it is better to collapse over secondary variables (even if they  
7503 are included in your statistical models) in order to control the den-  
7504 sity of the figure and draw attention to the key question of inter-  
7505 est.

## 7506 15.2 *Exploratory visualization*

7507 So far in this chapter we have focused on principles of *confirmatory* data  
7508 visualization: how to make production-quality figures that convey the  
7509 key preregistered analyses without hiding sources of variability or mis-  
7510 leading readers about the reliability of the results. Yet, this is only one  
7511 role that data visualization plays when doing science. An equally impor-  
7512 tant role is called *exploratory visualization*: the more routine practice of

7513 understanding one’s own data by visualizing it. This role is analogous to  
7514 the sense of exploratory data analyses discussed in chapter 11. We typi-  
7515 cally do not preregister exploratory visualizations, and when we decide  
7516 to include them in a paper they are typically in the service of a secondary  
7517 argument (e.g., checking the robustness of an effect or validating that  
7518 some assumption is satisfied).

7519 This kind of visualization plays a ubiquitous role in a researcher’s day-to-  
7520 day activities. While confirmatory visualization is primarily audience-  
7521 driven and concerned with visual communication, exploratory visual-  
7522 ization is first and foremost a “cognitive tool” for the researcher. The  
7523 first time we load in a new dataset, we start up a new feedback loop—  
7524 we ask ourselves questions and answer them by making visualizations.  
7525 These visualizations then raise further questions and are often our best  
7526 tool for debugging our code. In this section, we consider some best  
7527 practices for exploratory visualization.

### 7528 15.2.1 Examining distributional information

7529 The primary advantage of exploratory visualization—the reason it is  
7530 uniquely important for data science—is that it gives us access to holistic  
7531 information about the distribution of the data that cannot be captured  
7532 in any single summary statistic. The most famous example is known

as “Anscombe’s quartet,” a set of four datasets with identical statistics (figure 15.20). They have the same means, the same variances, the same correlation, the same regression line, and the same  $R^2$  value. Yet, when they are plotted, they reveal striking structural differences. The first looks like a noisy linear relationship—the kind of idealized relationship we imagine when we imagine a regression line. But the second is a perfect quadratic arc, the third is a perfectly noiseless line with a single outlier, and the fourth is nearly categorical: every observation except one shares exactly the same x-value.

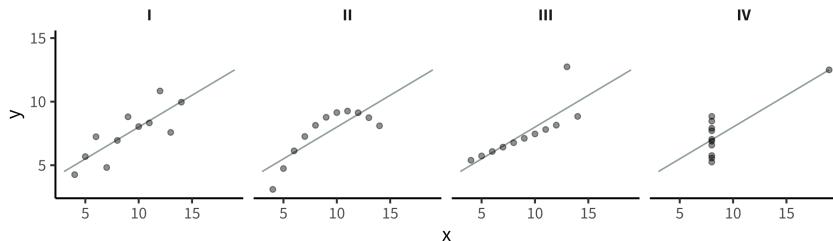


Figure 15.20  
Anscombe’s quartet (Anscombe 1973).

If our analyses are supposed to help us distinguish between different data-generating processes, corresponding to different psychological theories, it is clear that these four datasets would correspond to dramatically different theories even though they share the same statistics. Of course, there are arbitrarily many datasets with the same statistics, and most of these differences don’t matter (this is why they are called “summary” statistics, after all!). Figure 15.21 and table 15.1 show just how bad things can get when we rely on summary statistics. When we operationalize a theory’s predictions in terms of a single statistic (e.g., a dif-

7551 ference between groups or a regression coefficient), we can lose track  
7552 of everything else that may be going on. Good visualizations force us  
7553 to zoom out and take in the bigger picture.

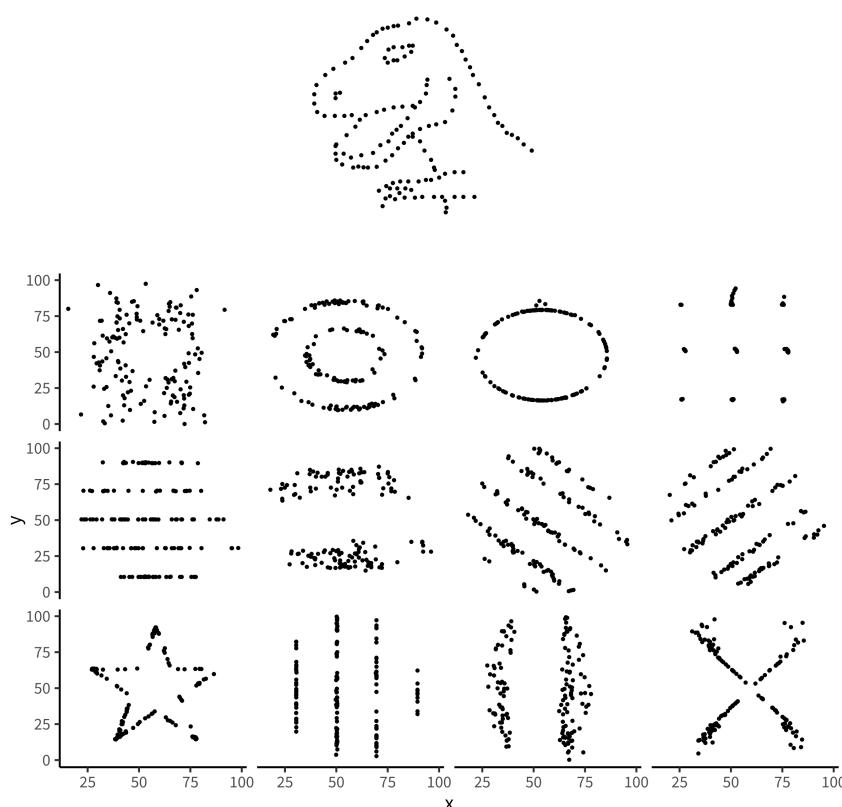


Figure 15.21

Originally inspired by the Datasaurus figure constructed by @albertocairo on Twitter using the DrawMyData tool (<http://robertgrantstats.co.uk/drawmydata.html>), we can construct an arbitrary number of different graphs with exactly the same statistics (Matejka and Fitzmaurice 2017; Murray and Wilson 2021), such as the Datasaurus Dozen (Matejka and Fitzmaurice 2017).

Table 15.1  
Summary statistics for each dataset in the Datasaurus Dozen (Matejka 2017).

dataset	mean_x	mean_y	sd_x	sd_y	cor_xy
away	54.3	47.8	16.8	26.9	-0.064
bullseye	54.3	47.8	16.8	26.9	-0.069
circle	54.3	47.8	16.8	26.9	-0.068
dino	54.3	47.8	16.8	26.9	-0.064
dots	54.3	47.8	16.8	26.9	-0.060
h_lines	54.3	47.8	16.8	26.9	-0.062
high_lines	54.3	47.8	16.8	26.9	-0.069

Table 15.1  
Summary statistics for each dataset in the Datasaurus Dozen (Matejka 2017).

dataset	mean_x	mean_y	sd_x	sd_y	cor_xy
slant_down	54.3	47.8	16.8	26.9	-0.069
slant_up	54.3	47.8	16.8	26.9	-0.069
star	54.3	47.8	16.8	26.9	-0.063
v_lines	54.3	47.8	16.8	26.9	-0.069
wide_lines	54.3	47.8	16.8	26.9	-0.067
x_shape	54.3	47.8	16.8	26.9	-0.066

## ⚠️ ACCIDENT REPORT

### *[Distributional] gorillas in our midst.*

Many data scientists don't bother checking what their data looks like before proceeding to test specific hypotheses. Yanai and Lercher (2020) cleverly designed an artificial dataset for their students to test for such oversight. Each row of the dataset contained an individual's body mass index (BMI) and the number of steps they walked on a given day. While the spreadsheet looked innocuous, the data was constructed such that simply plotting the raw data revealed a picture of a gorilla. One group of 19 students was given an explicit set of hypotheses to test (e.g., about the relationship between BMI and steps). Fourteen of these students failed to notice a gorilla, suggesting that they evaluated these hypotheses without ever visualizing their data. Another group of 14 students were simply asked what, if anything, they could conclude (without being given explicit hypotheses). More of these students apparently made the visualization, but five of them still failed to notice the gorilla (figure 15.22)!

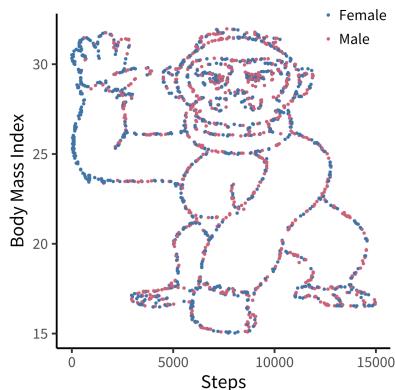


Figure 15.22

A dataset constructed by Yanai and Lercher (2020) that revealed a picture of a gorilla when the raw data were plotted.

While it may not be surprising that a group of students would take the shortest path to completing their assignment, similar concerns have been raised in much more serious cases concerning how experienced researchers could fail to notice obviously fraudulent data. For example, when the Datacolada bloggers (2021) made a simple histogram of the car mileage data reported in Shu et al. (2012; released publicly by Kristal et al. 2020), they were immediately able to observe that it followed a perfectly uniform distribution, truncated at exactly 50,000 miles (figure 15.23). Given a little thought, this pattern should be extremely puzzling. Over a given period of time, we would typically expect something more bell-shaped: a small number of people will drive very little (e.g., 1,000 miles), a small number of people will drive a lot (e.g., 50,000 miles), and most people will fall between these tails. So it is highly surprising to find exactly the same number of drivers in every mileage bin. While further specialized analyses revealed additional evidence of fraud (e.g., based on patterns of rounding and pairs of duplicated data points), this humble histogram was already enough to set off alarm bells. A recurring regret raised by the coauthors of this paper is that they never thought to make this visualization before reporting their statistical tests.

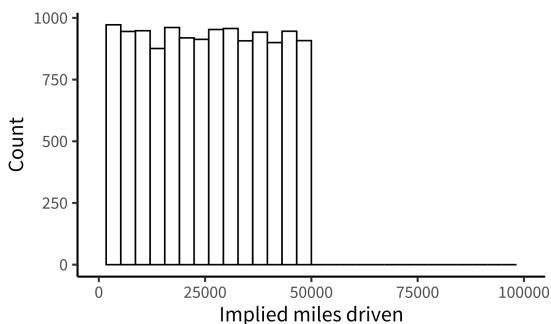


Figure 15.23

A suspiciously uniform distribution abruptly cutting off at 50K miles. Ring the alarm!

Our data are always messier than we expect. There might be a bug in our coding scheme, a column might be mislabeled, or it might contain a range of values that we didn't expect. Maybe our design wasn't perfectly balanced, or something went wrong with a particular participant's keyboard presses. Most of the time, it's not tractable to manually scroll through our raw data looking for such problems. Visualization is our first line of defense for the all-important process of running "data diagnostics." If there is a weird artifact in our data, it will pop out if we just make the right visualizations.

7556

### 7557 15.2.1 Data diagnostics

7558 So, which visualizations should we start with? The best practice is to  
7559 always start by making histograms of the raw data. As an example, let's  
7560 consider the rich and interesting dataset shared by Blake, McAuliffe,  
7561 and colleagues (2015) in their article "Ontogeny of Fairness in Seven

7562 Societies.” This article studies the emergence of children’s reasoning  
7563 about fairness—both when it benefits them and when it harms them—  
7564 across cultures.

 CODE

If you want to follow along with this example at home, you can load the data from our repository!

```
repo <- "https://raw.githubusercontent.com/langcog/experimentology/main"  
  
fairness_raw <- read_csv(file.path(repo, "data/viz/ontogeny_of_fairness.csv"))  
  
fairness <- fairness_raw |>  
  
  mutate(trial_num = trial |> str_remove("t") |> as.numeric(),  
         trial_type = eq.uneq |> fct_recode("Equal" = "E", "Unequal" = "U"),  
         condition = condition |> fct_recode("Advantageous" = "AI",  
                                               "Disadvantageous" = "DI"),  
         age = floor(actor.age.years),  
         reject = decision == "reject") |>  
  
  select(subj_id = actor.id, age, country, condition, trial_num, trial_type, reject) |>  
  arrange(country, condition, subj_id, trial_num)
```

7565

7566 In this study, pairs of children played the “inequity game”: they sat  
7567 across from one another and were given a particular allocation of snacks.  
7568 On some trials, each participant was allocated the same amount (“equal”  
7569 trials) and on some trials they were allocated different amounts (“un-  
7570 equal” trials). One participant was chosen to be the “actor” and got to  
7571 choose whether to accept or reject the allocation: in the case of rejec-  
7572 tion, neither participant got anything. The critical manipulation was

7573 between two forms of inequity. Some pairs were assigned to the “dis-  
7574 advantageous” condition, where the actor was allocated less than their  
7575 partner on unequal trials (e.g., one vs four). Others were assigned to the  
7576 “advantageous” condition, where they were allocated more (e.g., four  
7577 vs one).

7578 The confirmatory design plot for this study would focus on contrast-  
7579 ing developmental trajectories for advantageous vs disadvantageous in-  
7580 equality. However, this is a complex, multivariate dataset, including  
7581 866 pairs from different age groups and different testing sites across the  
7582 world which used subtly different protocols. How might we go about  
7583 the process of exploratory visualization for this dataset?

### 7584 15.2.2 Plot data collection details

7585 Let’s start by getting a handle on some of the basic sample character-  
7586 istics. For example, how many participants were in each age bin (fig-  
7587 ure 15.24)?

#### CODE

Exploratory histograms are often a combination of an aggregation step and a plotting step. In the aggregation step, we make use of the convenience `count()` function, which gives the number (`n`) of rows in a partic-

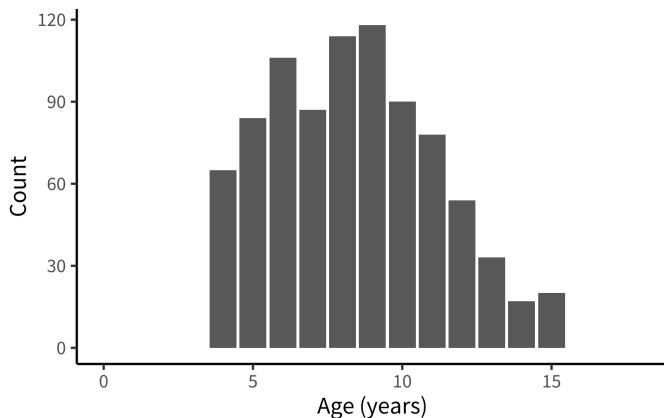


Figure 15.24  
Participants by age in the Blake data.

ular grouping. Here we count() twice in order to get first one row per participant and then count the number of participants within each age group.

```
fairness_by_age <- fairness |>
  count(age, subj_id) |>
  count(age)
```

And then we plot using ggplot():

```
ggplot(fairness_by_age, aes(x = age, y = n)) +
  geom_col() +
  xlim(0, 18) +
  labs(x = "Age (years)", y = "Count")
```

An alternative (perhaps more elegant) workflow here would be to use a histogram:

```
fairness_by_age <- fairness |>  
  count(age, subj_id)  
  
ggplot(fairness_by_age, aes(x = age)) +  
  geom_histogram(binwidth = 1) +  
  labs(x = "Age (years)", y = "Count")
```

Histograms are intended by ggplot to be for continuous data, however, and so they don't give the discrete bars that our earlier workflow did.

7591 How many participants were included from each country (fig-  
7592 ure 15.25)?

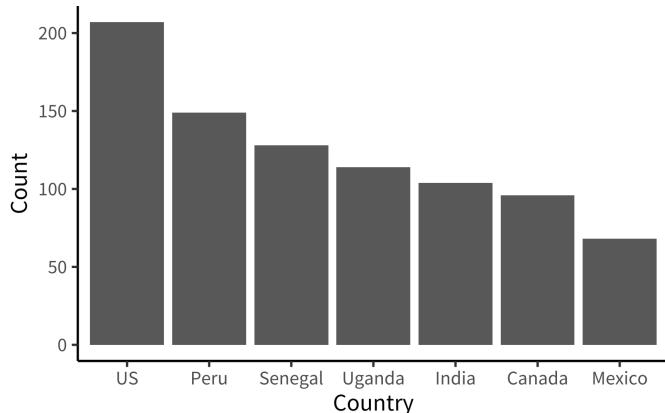


Figure 15.25  
Participants by country in the Blake data.

## CODE

Here we are going to make things even terser and use a pipe chain that *includes* the `ggplot()` call, just so we are writing only a single call to produce our plot. It's up to you whether you think this enhances the readability of your code or decreases it. We find that it's sometimes useful when you don't plan on keeping the intermediate data frame for any other use than plotting.

```
fairness |>  
  
  count(country, subj_id) |>  
  
  count(country) |>  
  
  mutate(country = fct_reorder(country, -n)) |>  
  
  ggplot(aes(x = country, y = n)) +  
  
    geom_col() +  
  
    labs(x = "Country", y = "Count")
```

If you use this technique, be careful to use pipe (`|>` or `%>%`) between function calls but use `(+)` between `ggplot` layers!

The only other trick to point out here is that we use the `fct_reorder()` call to order the levels of the `country` factor in descending order. This function is found in the very useful `forcats` package (Wickham 2023) of the `tidyverse`, which contains all sorts of functions for working with factors.

7595 Are ages roughly similar across each country (figure 15.26)?

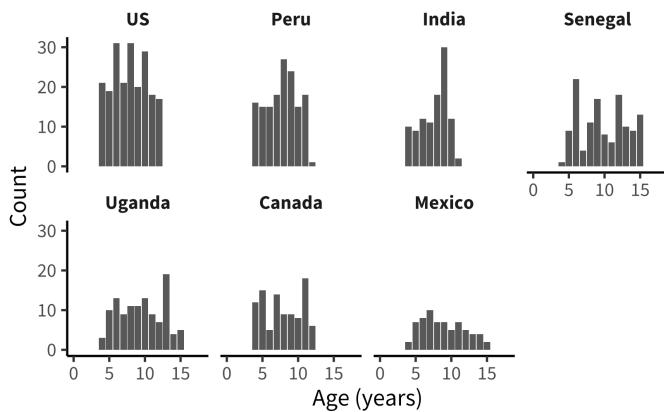


Figure 15.26  
Age distribution across countries in the Blake data.

## CODE

This next plot simply combines the grouping factors of each of the last two plots, and uses `facet_wrap()` to show a separate histogram by country:

```
fairness |>

  count(country, age, subj_id) |>

  count(country, age) |>

  mutate(country = fct_reorder(country, -n)) |>

  ggplot(aes(x = age, y = n)) +

    facet_wrap(vars(country), ncol = 4) +

    geom_col() +

    xlim(0, 18) +

    labs(x = "Age (years)", y = "Count")
```

7596

7597 These exploratory visualizations help us read off some descriptive prop-

7598    erties of the sample. For example, we can see that age ranges differ  
7599    somewhat across sites: the maximum age is 11 in India but 15 in Mex-  
7600    ico. We can also see that age groups are fairly imbalanced: in Canada,  
7601    there are 18 eleven-year-olds but only 5 six-year-olds.

7602    None of these properties are problematic, but seeing them gives us a  
7603    degree of awareness that could shape our downstream analytic decisions.  
7604    For example, if we did not appropriately model random effects, our  
7605    estimates would be dominated by the countries with larger sample sizes.  
7606    And if we were planning to compare specific groups of six-year-olds  
7607    (for some reason), this analysis would be underpowered.

### 7608    15.2.3 *Explorating distributions*

7609    Now that we have a handle on the sample, let's get a sense of the depen-  
7610    dent variable: the participant's decision to accept or reject the allocation.  
7611    Before we start taking means, let's look at how the "rejection rate" vari-  
7612    able is distributed. We'll aggregate at the participant level, and check  
7613    the frequency of different rejection rates, overall (figure 15.27).

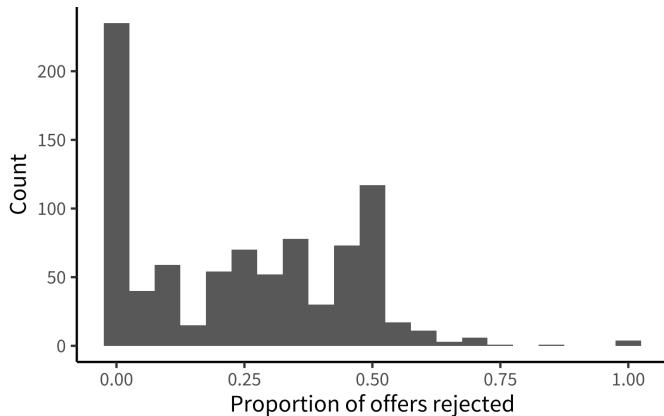


Figure 15.27  
Rejection rates in the Blake data.

## CODE

Rejection rate is a continuous variable, so we switch to using a histogram in this case, choosing 0.05 as a reasonable bin width to see the distribution.

```

fairness_by_subj <- fairness |>

filter(!is.na(trial_type)) |>

group_by(subj_id) |>

summarize(mean_reject = mean(reject, na.rm = TRUE))

ggplot(fairness_by_subj, aes(x = mean_reject)) +
  geom_histogram(binwidth = .05) +
  labs(x = "Proportion of offers rejected", y = "Count")

```

7614

7615 We notice that many participants (27%) never reject in the entire experiment. This kind of “zero-inflated” distribution is not uncommon in  
 7616 psychology, and may warrant special consideration when designing the  
 7617 statistical model. We also notice that there is clumping around certain

7619 values. This clumping leads us to check how many trials each participant  
 7620 is completing (figure 15.28).

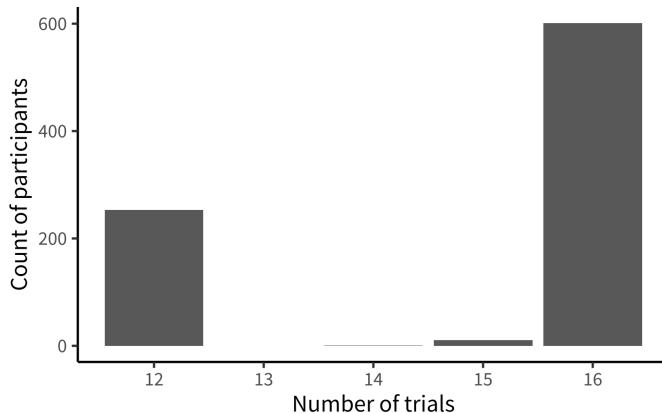


Figure 15.28  
 Trials per participant in the Blake data.

## CODE

This histogram is very similar to the ones above; however, we now use `count()` twice, first getting the trial counts for each participant and then counting how many times each count occurs overall!

```

fairness |>

  filter(!is.na(trial_type)) |>

  count(subj_id) |>

  count(n) |>

  ggplot(aes(x = n, y = nn)) +

  geom_col() +

  labs(x = "Number of trials", y = "Count of participants")

```

7621

7622 There's some variation here: most participants completed 17 trials, but  
 7623 some participants completed 8 trials, and a small number of participants

7624 have 14 or 15. Given the logistical complexity of large multi-site stud-  
7625 ies, it is common to have some changes in experimental protocol across  
7626 data collection. Indeed, looking at the supplement for the study, we  
7627 see that while India and Peru had 12 trials, additional trials were added  
7628 at the other sites. In a design where the number of trials was carefully  
7629 controlled, seeing unexpected numbers here (like the 14 or 15 trial bins)  
7630 are clues that something else may be going on in the data. In this case, it  
7631 was a small number of trials with missing data. More generally, seeing  
7632 this kind of signal in a visualization of our own data typically leads us  
7633 to look up the participant IDs in these bins and manually inspect their  
7634 data to see what might be going on.

#### 7635 15.2.4 *Hypothesis-driven exploration*

7636 Finally, we can make a few versions of the design plot that are broken  
7637 out by different variables. Let's start by just looking at the data from the  
7638 largest site (figure 15.29).

7639 Figure 15.29 is not a figure we'd put in a paper, but it helps us get a sense  
7640 of the pattern in the data. There appears to be an age trend that's specific  
7641 to the Unequal trials, with rejection rates rising over time (compared to  
7642 roughly even or decreasing rates in the Equal trials). Meanwhile, rejec-  
7643 tion rates for the disadvantaged group also seem slightly higher than

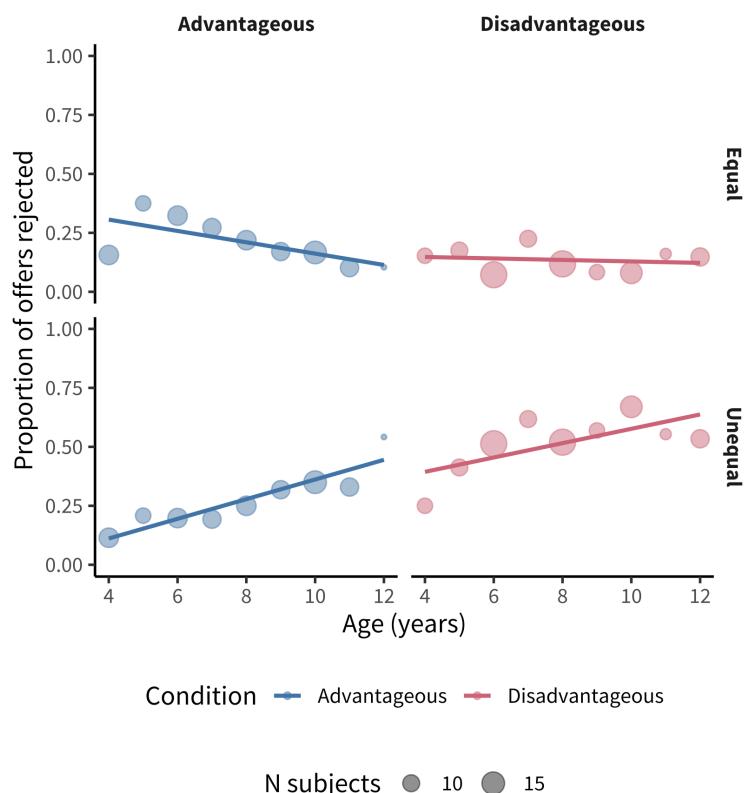


Figure 15.29  
Rejection rates in the US data from  
Blake, plotted by age.

7644 those in the advantageous group.

 CODE

Here, we are using `geom_smooth()` to overlay regression trends over the raw data. `geom_smooth()` takes a number of different options corresponding to different smoothing techniques. Nonparametric smoothing can be a good choice for exploratory visualizations if you have a lot of data and want to make minimal assumptions about the form of the trend.

Here, however, we show the linear regression trend, `geom_smooth(method = "lm")`, which better corresponds to the predictions of the study and the statistical model being used (see chapter 7). Other regression forms can be specified with the `formula` argument. For example, we could show quadratic smoothing with `geom_smooth(method = "lm", formula = y ~ poly(x, 2))`. The form of smoothing you use may differ across exploratory and confirmatory visualizations. In a confirmatory visualization—if you are going to include a smoothing curve—it is typically best to use the one specified by your statistical model, as the slopes will correspond to the inferences being tested.

We begin by making a summary dataset:

```
fairness_by_age <- fairness |>  
  filter(!is.na(reject)) |>  
  group_by(country, trial_type, condition, age, subj_id) |>  
  summarize(mean_reject_subj = mean(reject, na.rm = TRUE)) |>  
  group_by(country, trial_type, condition, age) |>  
  summarize(mean_reject_age = mean(mean_reject_subj, na.rm = TRUE),  
            n_subj = n()) |>  
  ungroup()
```

Then we can create the visualization:

```
fairness_by_age |> filter(country == "US") |>  
  ggplot(aes(x = age, y = mean_reject_age, color = condition)) +  
  facet_grid(vars(trial_type), vars(condition)) +  
  geom_smooth(method = "lm", se = FALSE) +  
  geom_point(aes(size = n_subj), alpha = .5) +  
  ylim(c(0, 1)) +  
  labs(x = "Age (years)", y = "Proportion of offers rejected",  
       color = "Condition", size = "N subjects") +  
  theme(legend.position = "bottom", legend.box = "vertical")
```

We often find it convenient to filter the summary dataset in the plotting call, so that we can reuse it again.

7646

Now let's rebin the data into two-year age groups so that individual

7648 point estimates are a bit more reliable, and add the other countries back

7649 in.<sup>7</sup>

<sup>7</sup> Binning data is a trick that we often use for reducing complexity in a plot when data are noisy. It should be used with care, however, since different binning decisions can sometimes lead to different conclusions. Here we tried several binning intervals and decided that two-year age bins showed the underlying trends pretty well.

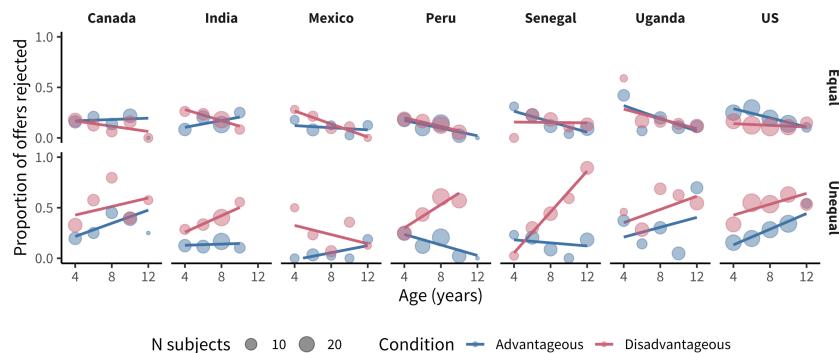


Figure 15.30  
Rejection rates by age for all data in the  
Blake dataset.

7650 Figure 15.30 is now looking much closer to a quick-and-dirty version  
 7651 of a “design plot” we might include in a paper. The DV (rejection rate)  
 7652 is on the y-axis, and the primary variable of interest (age) is on the x-  
 7653 axis. Other elements of the design (country and trial type) are mapped  
 7654 to color and facets, respectively.

### CODE

Despite the difference between the plot above and this one, the code to produce them is actually very similar. The only difference is the creation of the binned variable and a slight shift of aesthetic and faceting variables.

```
fairness_by_age_binned <- fairness |>  
  filter(!is.na(reject)) |>  
  mutate(age_binned = floor(age / 2) * 2) |>  
  group_by(country, trial_type, condition, age_binned, subj_id) |>  
  summarize(mean_reject_subj = mean(reject, na.rm = TRUE)) |>  
  group_by(country, trial_type, condition, age_binned) |>  
  summarize(mean_reject_age = mean(mean_reject_subj, na.rm = TRUE),  
            n = n()) |>  
  ungroup()  
  
ggplot(fairness_by_age_binned,  
       aes(x = age_binned, y = mean_reject_age, color = condition)) +  
  facet_grid(vars(trial_type), vars(country)) +  
  geom_smooth(method = "lm", se = FALSE, aes(weight=n)) +  
  geom_point(alpha = .5, aes(size = n)) +  
  scale_x_continuous(breaks = seq(4, 12, 4), limits = c(3,13)) +  
  scale_y_continuous(limits = c(0, 1), breaks = c(0, .5, 1)) +  
  labs(x = "Age (years)", y = "Proportion of offers rejected",  
        color = "Condition", size = "N subjects") +  
  theme(legend.position = "bottom")
```

7657 *15.2.5 Visualization as debugging*

7658 The point of exploratory visualization is to converge toward a better  
7659 understanding of what's going on in your data. As you iterate through  
7660 different exploratory visualizations, *stay vigilant!* Think about what you  
7661 expect to see before making the plot, then ask yourself whether you got  
7662 what you expected. You can think of this workflow as a form of "visual  
7663 debugging." You might notice a data point with an impossible value,  
7664 such as a proportion greater than one or a reaction time less than zero.  
7665 Or you might notice weird clusters or striations, which might indicate  
7666 heterogeneity in data entry (perhaps different coders used slightly dif-  
7667 ferent rubrics or rounded in different ways). You might notice that an  
7668 attribute is missing for some values and trace it back to a bug reading in  
7669 the data or merging data frames (maybe there was a missing comma in  
7670 our csv file). If you see anything that looks weird, track it down until  
7671 you understand why it's happening. Bugs that are subtle and invisible  
7672 in other parts of the analysis pipeline will often pop out as red flags in  
7673 visualizations.

7674 *15.3 Chapter summary: Visualization*

7675 This chapter has given a short review of the principles of data visual-  
7676 ization, especially focusing on the needs of experimental psychology,

7677 which are often quite different than those of other fields. We partic-  
7678 ularly focused on the need to make visualization part of the experi-  
7679 menter's analytic workflow. Picking up the idea of a "default model"  
7680 from chapter 7, we discussed a default "design plot" that reflects the  
7681 key choices made in the experimental design. Within this framework,  
7682 we then discussed different visualizations of distribution and variability  
7683 that better align our graphics with the principles of measurement and  
7684 attention to raw data that we have been advocating throughout.



### DISCUSSION QUESTIONS

1. Choose a recent piece of research that you've heard about and try to sketch the "design plot" with pencil and paper. What does and doesn't work? How does your sketch differ from the visualizations in the paper?
2. The "design plot" idea that we've discussed here can run into problems when an experimental design is too complex to show on a single plot. Imagine you had data from a trial of attention deficit hyperactivity disorder (ADHD) treatment that manipulated both whether a medication was given and whether patients received therapy in a crossed design. The researchers measured two different outcomes: parent report symptom severity and teacher report symptom severity in four different time-points (baseline, three months, six months, and nine months). How could you show the data from such an experiment in

a transparent way?

7686

## READINGS

There are many good introductions to data visualization. Here are two social-science focused books whose advice we agree with and that also contain a lot of practical information and helpful R code for the same packages we use here.

- Healy, Kieran (2018). *Data Visualization: A Practical Introduction*. Princeton University Press. Available free online at <https://socviz.co>.
- Wilke, Claus O. (2019). *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media. Available free online at <https://clauswilke.com/dataviz>.

For a more classical treatment, see:

- Tukey, John W. (1977). *Exploratory Data Analysis*. Pearson.
- Tufte, Edward R. (2001). *The Visual Display of Quantitative Information*. 2nd ed. Graphics Press.

7687

## 7688 References

- Allen, Micah, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Roger A. Kievit. 2019. “Raincloud Plots: A Multi-Platform Tool for Robust Data Visualization.” *Wellcome Open Research* 4.
- Anscombe, Francis J. 1973. “Graphs in Statistical Analysis.” *The American Statistician* 27 (1): 17–21.
- Arnold, Jeffrey B. 2023. *ggthemes: Extra Themes, Scales and Geoms for ggplot2*.  
7689 <https://CRAN.R-project.org/package=ggthemes>.
- Barnett, Samuel A., Thomas L. Griffiths, and Robert D. Hawkins. 2022. “A Pragmatic Account of the Weak Evidence Effect.” *Open Mind* 6 (2022): 169–182. [https://doi.org/10.1162/opmi\\_a\\_00061](https://doi.org/10.1162/opmi_a_00061).
- Blake, P. R., K. McAuliffe, J. Corbit, T. C. Callaghan, O. Barry, A. Bowie, L. Kleutsch, et al. 2015. “The Ontogeny of Fairness in Seven Societies.” *Nature* 528 (7581): 258–261.
- Börner, Katy, Andreas Bueckle, and Michael Ginda. 2019. “Data Visualization Literacy: Definitions, Conceptual Frameworks, Exercises, and Assessments.” *Proceedings of the National Academy of Sciences* 116 (6): 1857–1864.
- Braginsky, Mika, and Daniel Yurovsky. 2018. *tidyboot: Tidyverse-Compatible Bootstrapping*. <https://CRAN.R-project.org/package=tidyboot>.
- Brody, Howard, Michael Russell Rip, Peter Vinent-Johansen, Nigel Paneth, and Stephen Rachman. 2000. “Map-Making and Myth-Making in Broad Street: The London Cholera Epidemic, 1854.” *The Lancet* 356 (9223): 64–68.
- Cleveland, William S., and Robert McGill. 1984. “Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical

- Methods.” *Journal of the American Statistical Association* 79 (387): 531–554.
- Coppock, Alexander. 2019. “Visualize as You Randomize: Design-Based Statistical Graphs for Randomized Experiments.” In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Green, 320–336. Cambridge University Press.
- Datacolada. 2021. “Evidence of Fraud in an Influential Field Experiment about Dishonesty.” <https://datacolada.org/98>.
- Friendly, Michael, and Howard Wainer. 2021. *A History of Data Visualization and Graphic Communication*. Harvard University Press.
- Gelman, Andrew, and Antony Unwin. 2013. “Infovis and Statistical Graphics: Different Goals, Different Looks.” *Journal of Computational and Graphical Statistics* 22 (1): 2–28.
- Halliday, Stephen. 2001. “Death and Miasma in Victorian London: An Obstinate Belief.” *British Medical Journal* 323 (7327): 1469–1471.
- Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton University Press.
- Henry, Lionel, Hadley Wickham, and Winston Chang. 2022. *ggstance: Horizontal ggplot2 Components*. <https://CRAN.R-project.org/package=ggstance>.
- Hocking, Toby Dylan. 2023. *directlabels: Direct Labels for Multicolor Plots*. <https://CRAN.R-project.org/package=directlabels>.
- Kristal, Ariella S., Ashley V. Whillans, Max H. Bazerman, Francesca Gino, Lisa L. Shu, Nina Mazar, and Dan Ariely. 2020. “Signing at the Beginning Versus at the End Does Not Decrease Dishonesty.” *Proceedings of the National Academy of Sciences* 117 (13): 7103–7107.

- Mackinlay, Jock. 1986. "Automating the Design of Graphical Presentations of Relational Information." *ACM Transactions on Graphics* 5 (2): 110–141.
- Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–1294.
- Murray, Lori L., and John G. Wilson. 2021. "Generating Data Sets for Teaching the Importance of Regression Analysis." *Decision Sciences Journal of Innovative Education* 19 (2): 157–166.
- Roeder, Kathryn. 1994. "DNA Fingerprinting: A Review of the Controversy." *Statistical Science* 9 (2): 222–247.
- Shu, Lisa L., Nina Mazar, Francesca Gino, Dan Ariely, and Max H. Bazerman. 2012. "Signing at the Beginning Makes Ethics Salient and Decreases Dishonest Self-Reports in Comparison to Signing at the End." *Proceedings of the National Academy of Sciences* 109 (38): 15197–15200.
- Snow, John. 1854. "Dr. Snow's Report." In *Report on the Cholera Outbreak in the Parish of St. James, Westminster, During the Autumn of 1854*. John Churchill. <https://johnsnow.matrix.msu.edu/work.php?id=15-78-55>.
- Snow, John. 1855. *On the Mode of Communication of Cholera*. John Churchill.
- Stiller, Alex J., Noah D. Goodman, and Michael C. Frank. 2015. "Ad-Hoc Implicature in Preschool Children." *Language Learning and Development* 11 (2): 176–190.
- Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Graphics Press.
- Tukey, John W. 1977. *Exploratory Data Analysis*. Pearson.

Wainer, Howard. 1984. “How to Display Data Badly.” *The American Statistician* 38 (2): 137–147.

Wickham, Hadley. 2023. *forcats: Tools for Working with Categorical Variables (Factors)*. <https://CRAN.R-project.org/package=forcats>.

Wilke, Claus O. 2019. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O’Reilly Media.

Wilke, Claus O. 2023. *ggridges: Ridgeline Plots in ggplot2*. <https://CRAN.R-project.org/package=ggridges>.

Yanai, I., and M. Lercher. 2020. “A Hypothesis Is a Liability.” *Genome Biology* 21, 231 (2020). <https://doi.org/10.1186/s13059-020-02133-w>.

Zacks, Jeffrey M., and Steven L. Franconeri. 2020. “Designing Graphs for Decision-Makers.” *Policy Insights from the Behavioral and Brain Sciences* 7 (1): 52–63.

## 16 META-ANALYSIS

7694

### LEARNING GOALS

- Discuss the benefits of synthesizing evidence across studies
- Conduct a simple fixed-effects or random-effects meta-analysis
- Reason about the role of within-study and across-study biases in meta-analysis

7695

7696 Throughout this book, we have focused on how to design individual ex-

7697 periments that maximize MEASUREMENT PRECISION and minimize bias.

7698 But even when we do our best to get a precise, unbiased estimate in an

7699 individual experiment, one study can never be definitive. Variability in

7700 participant demographics, stimuli, and experimental methods may limit

7701 the GENERALIZABILITY of our findings. Additionally, even well-powered

7702 individual studies have some amount of statistical error, limiting their

7703 precision. Synthesizing evidence across studies is critical for developing

7704 a balanced and appropriately evolving view of the overall evidence on

7705 an effect of interest and for understanding sources of variation in the  
7706 effect.

7707 Synthesizing evidence rigorously takes more than putting a search term  
7708 into Google Scholar, downloading articles that look topical or inter-  
7709 esting, and qualitatively summarizing your impressions of those studies.

7710 While this ad hoc method can be an essential first step in performing  
7711 a literature review (Grant and Booth 2009), it is not systematic and  
7712 doesn't provide a *quantitative* summary of a particular effect. Further,  
7713 it doesn't tell you anything about potential biases in the literature—for  
7714 example, a bias for the publication of positive effects.

7715 To address these issues, a more systematic, quantitative review of the  
7716 literature is often more informative. This chapter focuses on a specific  
7717 type of quantitative review called **meta-analysis**: a method for combin-  
7718 ing effect sizes across different studies. (If you need a refresher on effect  
7719 size, see chapter 5, where we introduce the concept.)<sup>1</sup> We include a  
7720 chapter on meta-analysis in *Experimentology* because we believe it's an  
7721 important tool that can focus experimental researchers on issues of MEA-  
7722 SUREMENT PRECISION and BIAS REDUCTION, two of our key themes.

7723 By combining information from multiple studies, meta-analysis often  
7724 provides more precise estimates of an effect size than any single study.  
7725 In addition, meta-analysis also allows the researcher to look at the extent

<sup>1</sup> We'll primarily be using Cohen's  $d$ , the standardized difference between means, which we introduced in chapter 5. There are many more varieties of effect size available, but we focus here on  $d$  because it's common and easy to reason about in the context of the statistical tools we introduced in the earlier sections of the book.

7726 to which an effect varies across studies. If an effect does vary across stud-  
7727 ies, meta-analysis also can be used to test whether certain study charac-  
7728 teristics systematically produce different results (e.g., whether an effect  
7729 is larger in certain populations).



### CASE STUDY

#### *Towel reuse by hotel guests*

Imagine you are staying in a hotel and you have just taken a shower. Do you throw the towels on the floor or hang them back up again? In a widely cited study on the power of social norms, Goldstein, Cialdini, and Griskevicius (2008) manipulated whether a sign encouraging guests to reuse towels focused on environmental impacts (e.g., “help reduce water use”) or social norms (e.g., “most guests reuse their towels”). Across two studies, they found that guests were significantly more likely to reuse their towels after receiving the social norm message (Study 1: odds ratio [OR] = 1.46, 95% CI [1.00, 2.16],  $p = 0.05$ ; Study 2: OR = 1.35, 95% CI [1.04, 1.77],  $p = 0.03$ ).

However, five subsequent studies by other researchers did not find significant evidence that social norm messaging increased towel reuse. (ORs ranged from 0.22 to 1.34, and no hypothesis-consistent  $p$ -value was less than 0.05). This caused many researchers to wonder if there is any effect at all. To examine this question, Scheibehenne, Jamil, and Wagenmakers (2016) statistically combined evidence across the studies via meta-analysis. This meta-analysis indicated that using social norm messages did signifi-

cantly increase hotel towel reuse, on average ( $OR = 1.26$ , 95% CI [1.07, 1.46],  $p < 0.005$ ). This case study demonstrates an important strength of meta-analysis: by pooling evidence from multiple studies, meta-analysis can generate more powerful insights than any one study alone. We will also see how meta-analysis can be used to assess variability in effects across studies.

7731

7732 Meta-analysis often teaches us something about a body of evidence that  
7733 we do not intuitively grasp when we casually read through a bunch of  
7734 articles. In the above case study, merely reading the individual studies  
7735 might give the impression that social norm messages do not increase  
7736 hotel towel reuse. But meta-analysis indicated that the average effect is  
7737 beneficial, although there might be substantial variation in effect sizes  
7738 across studies.<sup>2</sup>

### 7739 *16.1 The basics of evidence synthesis*

7740 As we explore the details of conducting a meta-analysis, we'll turn to  
7741 another running example: a meta-analysis of studies investigating the  
7742 "contact hypothesis" on intergroup relations.

7743 According to the contact hypothesis, prejudice toward members of mi-  
7744 nority groups can be reduced through intergroup contact interventions,

7745 in which members of majority and minority groups work together to  
 7746 pursue a common goal (Allport, Clark, and Pettigrew 1954). To ag-  
 7747 gregate the evidence on the contact hypothesis, Paluck, Green, and  
 7748 Green (2019) meta-analyzed studies that tested the effects of random-  
 7749 ized intergroup contact interventions on long-term prejudice-related  
 7750 outcomes.

7751 Using a systematic literature search, Paluck, Green, and Green (2019)  
 7752 searched for all papers that tested these effects and then extracted effect  
 7753 size estimates from each paper.<sup>3</sup> Because not every paper reports stan-  
 7754 dardized effect sizes—or even means and standard deviations for every  
 7755 group—this process can often involve scraping information from plots,  
 7756 tables, and statistical tests to try to reconstruct effect sizes.<sup>4</sup>

7757 Following best practices for meta-analysis (where there are almost never  
 7758 privacy concerns to worry about), Paluck, Green, and Green (2019)  
 7759 shared their data openly. The first few lines are shown in table 16.1.  
 7760 We'll use these data as our running example throughout.

Table 16.1

The first few lines of extracted effect sizes ( $d$ ) and their variances ( $\text{var}_d$ ) in the Paluck, Green, and Green (2019) meta-analysis.

name	pub_date	target	n_total	d	var_d
Boisjoly 06 B	2006	race	1243	0.030	0.006
Sorensen 10	2010	race	597	0.302	0.007

<sup>3</sup> This book will not cover the process of conducting a systematic literature search and extracting effect sizes, but these topics are critical to understand if you plan to conduct a meta-analysis or other evidence synthesis. Our experience is that extracting effect sizes from papers with inconsistent reporting standards can be especially tricky, so it can be helpful to talk to someone with experience in meta-analysis to get advice about this.

<sup>4</sup> For example, if the outcome variable is continuous, we could estimate Cohen's  $d$  from group means and standard deviations reported in the paper, even without having access to raw data.

Table 16.1

The first few lines of extracted effect sizes ( $d$ ) and their variances ( $\text{var}_d$ ) in the Paluck, Green, and Green (2019) meta-analysis.

name	pub_date	target	n_total	d	var_d
Scacco 18	2018	religion	474	0.000	0.010
Finseraas 2017	2017	foreigners	577	0.000	0.011
Sheare 74	1974	disability	400	1.059	0.011
Barnhardt 09	2009	religion	312	0.395	0.015

<sup>7761</sup> As we've seen throughout this book, visualizing data before and after  
<sup>7762</sup> analysis helps benchmark and check our intuitions about the formal sta-  
<sup>7763</sup> tistical results. In a meta-analysis, a common way to plot effect sizes  
<sup>7764</sup> is the **forest plot**, which depicts individual studies' estimates and con-  
<sup>7765</sup> fidence intervals. In the forest plot in figure 16.1,<sup>5</sup> the larger squares  
<sup>7766</sup> correspond to more precise studies; notice how much narrower their  
<sup>7767</sup> confidence intervals are than the confidence intervals of less precise stud-  
<sup>7768</sup> ies.

<sup>5</sup> You can ignore for now the final line, "RE Model"; we will return to this later.

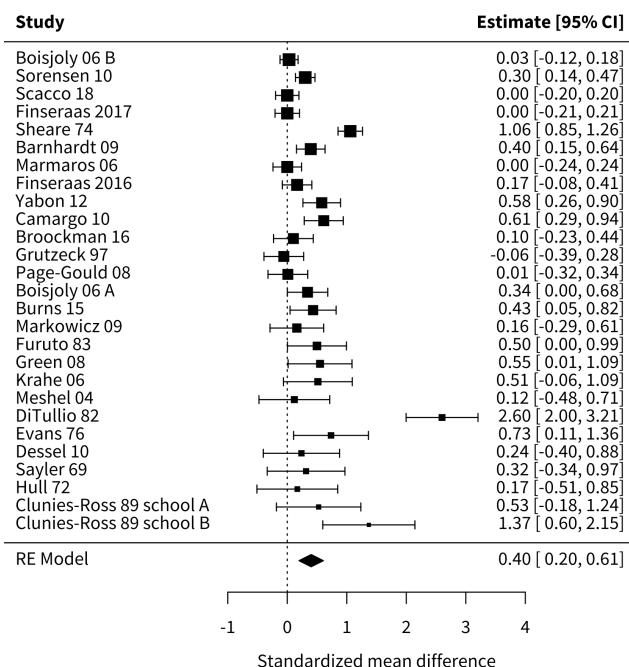


Figure 16.1

A forest plot for Paluck, Green, and Green (2019) meta-analysis. Studies are ordered from smallest to largest standard error.

## CODE

In this chapter, we use the wonderful `metafor` package (Viechtbauer 2010). With this package, you must first fit your meta-analytic model. But once you've fit your model `mod`, you can simply call `forest(mod)` to create a plot like the one above.

7769

### 16.1.1 How not to synthesize evidence

Many people's first instinct in evidence synthesis is to count how many studies supported versus did not support the hypothesis under investigation. This technique usually amounts to counting the num-

7770

7771

7772

7773

ber of studies with “significant”  $p$ -values, since—for better or for worse—“significance” is largely what drives the take-home conclusions researchers report (McShane and Gal 2017; Nelson, Rosenthal, and Rosnow 1986). In meta-analysis, we call this practice of counting the number of significant  $p$ -values **vote-counting** (Borenstein et al. 2021). For example, in the Paluck, Green, and Green (2019) meta-analysis, almost all studies had a positive effect size, but only 12 of 27 were significant. So, based on this vote-count, we would have the impression that most studies do not support the contact hypothesis.

Many qualitative literature reviews use this vote-counting approach, although often not explicitly. Despite its intuitive appeal, vote-counting can be very misleading because it characterizes evidence solely in terms of dichotomized  $p$ -values, while entirely ignoring effect sizes. In chapter 3, we saw how fetishizing statistical significance can mislead us when we consider individual studies. These problems also apply when considering multiple studies.

For example, small studies may consistently produce nonsignificant effects due to their limited power. But when many such studies are combined in a meta-analysis, the meta-analysis may provide strong evidence of a positive average effect. Inversely, many studies might have statistically significant effects, but if their effect sizes are small, then a meta-

7795 analysis might indicate that the average effect size is too small to be  
7796 practically meaningful. In these cases, vote-counting based on statistical  
7797 significance can lead us badly astray (Borenstein et al. 2021). To avoid  
7798 these pitfalls, meta-analysis combines the effect size estimates from each  
7799 study (not just their  $p$ -values), weighting them in a principled way.

7800 *16.1.2 Fixed-effects meta-analysis*

7801 If vote-counting is a bad idea, how should we combine results across  
7802 studies? Another intuitive approach might be to average effect sizes  
7803 from each study. For example, in Paluck et al.’s meta-analysis, the mean  
7804 of the studies’ effect size estimates is 0.44. This averaging approach is a  
7805 step in the right direction, but it has an important limitation: averaging  
7806 effect size estimates gives equal weight to each study. A small study  
7807 (e.g., Clunies-Ross and O’Meara 1989 with  $N = 30$ ) contributes as  
7808 much to the mean effect size as a large study (e.g., Boisjoly et al. 2006  
7809 with  $N = 1,243$ ). Larger studies provide more precise estimates of  
7810 effect sizes than small studies, so weighting all studies equally is not ideal.  
7811 Instead, larger studies should carry more weight in the analysis.

7812 To address this issue, **fixed-effects meta-analysis** uses a **weighted aver-**  
7813 **age** approach. Larger, more precise studies are given more weight in the  
7814 calculation of the overall effect size. Specifically, each study is weighted

7815 by the inverse of its variance (i.e., the inverse of its squared standard er-  
 7816 ror). This makes sense because larger, more precise studies have smaller  
 7817 variances, and thus get more weight in the analysis.

7818 In general terms, the fixed-effect pooled estimate is:

$$\hat{\mu} = \frac{\sum_{i=1}^k w_i \hat{\theta}_i}{\sum_{i=1}^k w_i}$$

7819 where  $k$  is the number of studies,  $\hat{\theta}_i$  is the point estimate of the  $i^{th}$  study,  
 7820 and  $w_i = 1/\hat{\sigma}_i^2$  is study  $i$ 's weight in the analysis (i.e., the inverse of its  
 7821 variance).<sup>6</sup>

7822 Using the fixed-effects formula, we can estimate that the overall effect  
 7823 size in Paluck et al.'s meta-analysis is a standardized mean difference  
 7824 of  $\hat{\mu} = 0.28$ ; 95% confidence interval [0.23, 0.34];  $p < 0.001$ . Because  
 7825 Cohen's  $d$  is our effect size index, this estimate would suggest that in-  
 7826 tergroup contact decreased prejudice by 0.28 standard deviations.

<sup>6</sup> If you are curious, the standard error of the fixed-effect  $\hat{\mu}$  is  $\frac{1}{\sum_{i=1}^k w_i}$ . This standard error can be used to construct a confidence interval or  $p$ -value, as described in chapter 6.

## CODE

Fitting meta-analytic models in `metafor` is quite simple. For example, for the fixed-effects model above, we simply ran the `rma()` function and specified that we wanted a fixed-effects analysis.

```
fe_model <- rma(yi = d, vi = var_d, data = paluck, method = "FE")
```

Then `summary(fe_model)` gives us the relevant information about the

fitted model.

7828

### 7829 16.1.3 Limitations of fixed-effects meta-analysis

7830 One of the limitations of fixed-effect meta-analysis is that it assumes  
7831 that the true effect size is, well, *fixed*! In other words, fixed-effect meta-  
7832 analysis assumes that there is a single effect size that all studies are es-  
7833 timating. This is a stringent assumption. It's easy to imagine that it  
7834 could be violated. Imagine, for example, that intergroup contact de-  
7835 creased prejudice when the group succeeded at its joint goal but *increased*  
7836 prejudice when the group failed. If we meta-analyzed two studies un-  
7837 der these conditions—one in which intergroup contact substantially in-  
7838 creased prejudice and one in which intergroup contact substantially de-  
7839 creased prejudice—it might appear that the true effect of intergroup  
7840 contact was close to zero, when in fact both of the meta-analyzed stud-  
7841 ies had large effects.

7842 In Paluck et al.'s meta-analysis, studies differed in several ways that  
7843 could lead to different true effects. For example, some studies recruited  
7844 adult participants while others recruited children. If intergroup contact  
7845 is more or less effective for adults versus children, then it is misleading  
7846 to talk about a single (i.e., "fixed") intergroup contact effect. Instead,

7847 we would say that the effects of intergroup contact vary across studies,

7848 an idea called **heterogeneity**.

7849 Does the concept of heterogeneity remind you of anything from when

7850 we analyzed repeated-measures data in chapter 7 on models? Recall

7851 that, with repeated-measures data, we had to deal with the possibility

7852 of heterogeneity across participants—and of the ways we did so was by

7853 introducing participant-level random intercepts to our regression model.

7854 It turns out that we can do a similar thing in meta-analysis to deal with

7855 heterogeneity across studies.

#### 7856 16.1.4 Random-effects meta-analysis

7857 While fixed-effect meta-analysis essentially assumes that all studies

7858 in the meta-analysis have the same population effect size,  $\mu$ , random-

7859 effects meta-analysis instead assumes that study effects come from

7860 a normal distribution with mean  $\mu$  and standard deviation  $\tau$ .<sup>7</sup> The

7861 larger the standard deviation,  $\tau$ , the more heterogeneous the effects are

7862 across studies. A random-effects model then estimates both  $\mu$  and  $\tau$ ,

7863 for example by maximum likelihood (DerSimonian and Laird 1986;

7864 Brockwell and Gordon 2001).

7865 Like fixed-effect meta-analysis, the random-effects estimate of  $\hat{\mu}$  is still

<sup>7</sup> Technically, other specifications of random-effects meta-analysis are possible. For example, robust variance estimation does not require making assumptions about the distribution of effects across studies (Hedges, Tipton, and Johnson 2010). These approaches also have other substantial advantages, like their ability to handle effects that are clustered, e.g., because some papers contribute multiple estimates (Hedges, Tipton, and Johnson 2010; Pustejovsky and Tipton 2021), and their ability to provide better inference in meta-analyses with relatively few studies (Tipton 2015). For these reasons, we often use these robust methods.

7866 a weighted average of studies' effect size estimates:

$$\hat{\mu} = \frac{\sum_{i=1}^k w_i \hat{\theta}_i}{\sum_{i=1}^k w_i}$$

7867 However, in random-effects meta-analysis, the inverse-variance

7868 weights now incorporate heterogeneity:  $w_i = 1 / (\hat{\tau}^2 + \hat{\sigma}_i^2)$ . Where

7869 before we had one term in our weights, now we have two. That

7870 is because these weights represent the inverse of studies' *marginal*

7871 variances, taking into account both statistical error due to their finite

7872 sample sizes ( $\hat{\sigma}_i^2$  as before) and also genuine effect heterogeneity ( $\hat{\tau}^2$ ).

7873 Conducting a random-effects meta-analysis of Paluck et al.'s dataset

7874 yields  $\hat{\mu} = 0.4$ ; 95% confidence interval [0.2, 0.61];  $p < 0.001$ . That

7875 is, *on average across studies*, intergroup contact was associated with a de-

7876 crease in prejudice of 0.4 standard deviations, substantially larger than

7877 the estimate from the fixed-effects model. This meta-analytic estimate

7878 is shown as the bottom line of figure 16.1.

## CODE

Fitting a random-effects model requires only a small change to the methods argument of `rma()`. (We also include the `knha` flag that adds a correction to the computation of standard errors and p-values.)

```
re_model <- rma(yi = d, vi = var_d, data = paluck, method = "REML", knha = TRUE)
```

7880 Based on the random-effects model, intergroup contact effects appear  
 7881 to differ across studies. Paluck et al. estimated that the standard devi-  
 7882 ation of effects across studies was  $\hat{\tau} = 0.44$ ; 95% confidence interval  
 7883 [0.25, 0.57]. This estimate indicates a substantial amount of heterogene-  
 7884 ity! To visualize these results, we can plot the estimated density of the  
 7885 population effects, which is just a normal distribution with mean  $\hat{\mu}$  and  
 7886 standard deviation  $\hat{\tau}$  (figure 16.2).

7887 This meta-analysis highlights an important point: that the overall effect  
 7888 size estimate  $\hat{\mu}$  represents only the *mean* population effect across studies.  
 7889 It tells us nothing about how much the effects *vary* across studies. Thus,  
 7890 we recommend always reporting the heterogeneity estimate  $\hat{\tau}$ , prefer-  
 7891 ably along with other related metrics that help summarize the distribu-  
 7892 tion of effect sizes across studies (Riley, Higgins, and Deeks 2011; Wang  
 7893 and Lee 2019; Mathur and VanderWeele 2019, 2020a). Reporting the  
 7894 heterogeneity helps readers know how consistent or inconsistent the ef-  
 7895 fects are across studies, which may point to the need to investigate *mod-  
 7896 erators* of the effect (i.e., factors that are associated with larger or smaller  
 7897 effects, such as whether participants were adults or children).<sup>8</sup>

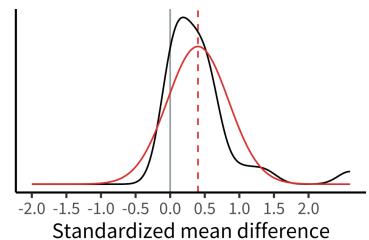


Figure 16.2  
 Estimated distribution of population ef-  
 fects from random-effects meta-analysis  
 of Paluck et. al's dataset (heavy red  
 curve) and estimated density of studies'  
 point estimates (thin black curve).

<sup>8</sup> One common approach to investi-  
 gating moderators in meta-analysis is  
 meta-regression, in which moderators  
 are included as covariates in a random-  
 effects meta-analysis model (Thompson  
 and Higgins 2002). As in standard re-  
 gression, coefficients can then be esti-  
 mated for each moderator, representing  
 the mean difference in population effect  
 between studies with versus without the  
 moderator.

 DEPTH

### *Single-paper meta-analysis and pooled analysis*

Thus far, we have described meta-analysis as a tool for summarizing results reported across multiple papers. However, some people have argued that meta-analysis should also be used to summarize the results of multiple studies reported in a single paper (Goh, Hall, and Rosenthal 2016).

For instance, in a paper where you describe three different experiments on a hypothesis, you could (1) extract summary information (e.g., means and standard deviations) from each study, (2) compute the effect size, and then (3) combine the effect sizes in a meta-analysis.

Single-paper meta-analyses come with many of the same strengths and weaknesses we have discussed thus far. One unique weakness, though, is that having a small number of studies means that you typically have low power to detect heterogeneity and moderators. This lack of power sometimes leads researchers to claim that there are no significant differences between their studies. But an alternative explanation is that there simply wasn't enough power to detect those differences!

As an alternative, you can also pool the actual data from the three studies, as opposed to just pooling summary statistics. For example, if you have data from 10 participants in each of the three experiments, you could pool them into a single dataset with 30 participants and include random effects of your condition manipulation across experiments (as described in chapter 7). This strategy is often referred to as **pooled** or **integrative**

data analysis (and occasionally as “mega-analysis,” which sounds cool).

Study 1			Pooled data analysis			Meta-analysis								
Study	Participant	Group	Prejudice	Age	Study	Participant	Group	Prejudice	Age	Study	Effect size ( $d$ )	Age		
1	1	Treatment	2	18	1	1	Treatment	2	18	1	8	18		
1	2	Treatment	2	18	1	2	Treatment	2	18	2	5	24		
1	3	Treatment	2	18	1	3	Treatment	2	18	3	1	45		
1	4	Treatment	2	18	1	4	Treatment	2	18					
1	5	Treatment	2	18	1	5	Treatment	2	18					
1	6	Control	10	18	1	6	Control	10	18					
1	7	Control	10	18	1	7	Control	10	18					
1	8	Control	10	18	1	8	Control	10	18					
1	9	Control	10	18	1	9	Control	10	18					
1	10	Control	10	18	1	10	Control	10	18					
Study 2			Pooled data analysis			Meta-analysis								
Study	Participant	Group	Prejudice	Age	Study	Participant	Group	Prejudice	Age	Study	Effect size ( $d$ )	Age		
2	1	Treatment	5	24	2	2	Treatment	5	24					
2	2	Treatment	5	24	2	4	Treatment	5	24					
2	3	Treatment	5	24	2	5	Treatment	5	24					
2	4	Treatment	5	24	2	6	Control	10	24					
2	5	Treatment	5	24	2	7	Control	10	24					
2	6	Control	10	24	2	8	Control	10	24					
2	7	Control	10	24	2	9	Control	10	24					
2	8	Control	10	24	2	10	Control	10	24					
2	9	Control	10	24	3	1	Treatment	9	45					
2	10	Control	10	24	3	2	Treatment	9	45					
Study 3			Pooled data analysis			Meta-analysis								
Study	Participant	Group	Prejudice	Age	Study	Participant	Group	Prejudice	Age	Study	Effect size ( $d$ )	Age		
3	1	Treatment	9	45	3	5	Treatment	9	45					
3	2	Treatment	9	45	3	6	Control	10	45					
3	3	Treatment	9	45	3	7	Control	10	45					
3	4	Treatment	9	45	3	8	Control	10	45					
3	5	Treatment	9	45	3	9	Control	10	45					
3	6	Control	10	45	3	10	Control	10	45					
3	7	Control	10	45										
3	8	Control	10	45										
3	9	Control	10	45										
3	10	Control	10	45										

Figure 16.3  
Meta-analysis vs pooled data analysis.

One of the benefits of pooled data analysis is that it can give you more power to detect moderators. For instance, imagine that the effect of an intergroup contact treatment is moderated by age. If we performed a traditional meta-analysis, we would only have three observations in our data set, yielding very low power. However, we have many more observations (and much more variation in the moderator) in the pooled data analysis, which can lead to higher power (figure 16.3).

Pooled data analysis is not without its own limitations (Cooper and Patall 2009). And, of course, sometimes it doesn't make as much sense to pool datasets (e.g., when measures are different from one another). Nonetheless, we believe that pooled data analysis and meta-analysis are both useful tools to keep in mind in a paper reporting multiple experiments!

7900 *16.2 Bias in meta-analysis*

7901 Meta-analysis is a great tool for synthesizing evidence across studies, but  
7902 the accuracy of a meta-analysis can be compromised by bias. We'll talk  
7903 about two categories of bias here: **within-study** and **across-study** biases.

7904 Either type can lead to meta-analytic estimates that are too large, too  
7905 small, or even in the wrong direction altogether.

7906 *16.2.1 Within-study biases*

7907 Within-study biases—such as demand characteristics, confounds, and  
7908 order effects, all discussed in chapter 9—not only impact the validity of  
7909 individual studies but also any attempt to synthesize those studies. And  
7910 of course, if individual study results are affected by analytic flexibility  
7911 (*p*-hacking), meta-analyzing these will result in inflated effect size esti-  
7912 mates. In other words: garbage in, garbage out.

7913 For example, Paluck, Green, and Green (2019) noted that early stud-  
7914 ies on intergroup contact almost exclusively used nonrandomized de-  
7915 signs. Imagine a hypothetical study where researchers studied a com-  
7916 pletely ineffective intergroup contact intervention, and nonrandomly  
7917 assigned low-prejudice people to the intergroup contact condition and  
7918 high-prejudice people to the control condition. In a scenario like this,  
7919 the researcher would of course find that the prejudice was lower in the

7920 intergroup contact condition. But the effect would not be a true contact  
7921 intervention effect, but rather a spurious effect of nonrandom assign-  
7922 ment (i.e., confounding). Now imagine meta-analyzing many studies  
7923 with similarly poor designs. The meta-analyst might find impressive  
7924 evidence of an intergroup contact effect, even if none existed.

7925 To mitigate this problem, meta-analysts often exclude studies that  
7926 may be especially affected by within-study bias. (For example, Paluck,  
7927 Green, and Green 2019 excluded nonrandomized studies). Of course,  
7928 these decisions can't be made on the basis of their effects on the  
7929 meta-analytic estimate or else this post hoc exclusion itself will lead to  
7930 bias! For this reason, inclusion and exclusion criteria for meta-analyses  
7931 should be preregistered whenever possible.

7932 Sometimes certain sources of bias cannot be eliminated by excluding  
7933 studies—often because studies in a particular domain share certain fun-  
7934 damental limitations (for example, attrition in drug trials). After data  
7935 have been collected, meta-analysts should also assess studies' risks of bias  
7936 qualitatively using established rating tools (Sterne et al. 2016). Doing so  
7937 allows the meta-analyst to communicate how much within-study bias  
7938 there may be.<sup>9</sup>

7939 Meta-analysts can also conduct sensitivity analyses to assess how much  
7940 results might be affected by different within-study biases or by exclud-

<sup>9</sup> If you're interested in assessing within-study bias, you can take a look at the Risk of Bias tool (<https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool>) developed by Cochrane, an organization devoted to evidence synthesis.

ing certain types of studies (Mathur and VanderWeele 2022). For example, if nonrandom assignment is a concern, a meta-analyst may run the analyses including only randomized studies, versus including all studies, in order to determine how much including nonrandomized studies changes the meta-analytic estimate. These two options parallel our discussion of experimental preregistration in chapter 11: to allay concerns about results-dependent meta-analysis, researchers can either preregister their analyses ahead of time or else be transparent about their choices after the fact. Sensitivity analyses can allay concerns that a specific choice of exclusion criteria is critically related to the reported results.

### 16.2.2 Across-study biases

Across-study biases occur if, for example, researchers selectively report certain types of findings or selectively publish certain types of findings (publication bias, as discussed in chapter 3 and chapter 11). Often, these across-study biases favor statistically significant positive results, which means the meta-analytic estimate based on those studies will be inflated relative to the true effect.

 ACCIDENT REPORT

## *Quantifying publication bias in the social sciences*

It's typically very hard to quantify publication bias because you don't know how many studies are out there in researchers' "file drawers"—unpublished studies are by definition not available. But a recent study took advantage of a unique opportunity to try and quantify publication bias within a known pool of studies.

Time-sharing Experiments in the Social Sciences (TESS) is an innovative project that lets researchers apply to run experiments on nationally representative samples in the US. In 2014, Franco, Malhotra, and Simonovits (2014) and colleagues took advantage of this application process by examining the entire population of 221 studies conducted through TESS.

Using this information, Franco and colleagues examined the records of these studies to determine whether the researchers found statistically significant results, a mixture of statistically significant and nonsignificant results, or only nonsignificant results. Then, they examined the likelihood that these results were published in the scientific literature.

Over 60% of studies with statistically significant results were published, compared to a mere 25% of studies that produced only statistically non-significant results. This finding was important because it quantified how strong publication bias actually was, at least in one particular population of studies. This estimate may not be general: for example, perhaps TESS studies were easier to put in the file drawer because they cost nothing for

the researchers to run. But even a lower level of publication bias can have a substantial effect on a meta-analysis, meaning that it is crucial to check for—and potentially, correct for—publication bias.

7960

7961 Like within-study biases, meta-analysts often try to mitigate across-  
7962 study biases by being careful about what studies make it into the  
7963 meta-analysis. Meta-analysts don't only want to capture high-profile,  
7964 published studies on their effect of interest but also studies published in  
7965 low-profile journals and the so-called gray literature (i.e., unpublished  
7966 dissertations and theses; Lefebvre et al. 2019).<sup>10</sup>

7967 There are also statistical methods to help assess how robust the results  
7968 may be to across-study biases. Among the most popular tools to assess  
7969 and correct for publication bias is the funnel plot (Duval and Tweedie  
7970 2000; Egger et al. 1997). A funnel plot shows the relationship between  
7971 studies' effect estimates and their precision (usually their standard error).  
7972 These plots are called “funnel plots” because, if there is no publication  
7973 bias, then as precision increases, the effects “funnel” toward the meta-  
7974 analytic estimate. As the precision is smaller, they spread out more be-  
7975 cause of greater measurement error. Figure 16.4] is an example of one  
7976 type of funnel plot (Mathur and VanderWeele 2020b) for a simulated  
7977 meta-analysis of 100 studies with no publication bias.

Tsuji et al. 2020 Mathur and VanderWeele 2021

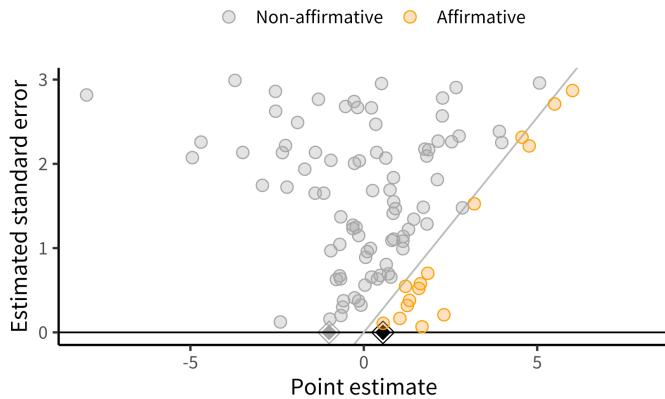


Figure 16.4

A significance funnel plot for a meta-analysis simulated to have no publication bias. Orange points: studies with  $p < 0.05$  and positive estimates. Grey points: studies with  $p \geq 0.05$  or negative estimates. Black diamond: random-effects estimate of  $\hat{\mu}$ .

## CODE

For this plot, we use the `PublicationBias` package (Braginsky, Mathur, and VanderWeele 2023) and the `significance_funnel()` function. (An alternative function is the `metafor` function `funnel()`, which results in a more “classic” funnel plot.) We use our fitted model `re_model`:

```
significance_funnel(yi = re_model$yi, vi = re_model$vi)
```

Because meta-analysis is such a well-established method, many of the relevant operations are “plug and play.”

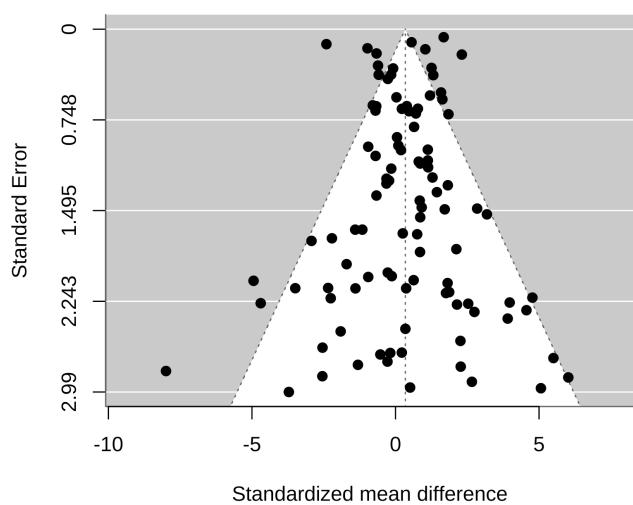


Figure 16.5  
A classic funnel plot.

<sup>7979</sup> As implied by the “funnel” moniker, our plot looks a little like a funnel.

<sup>7980</sup> Larger studies (those with smaller standard errors) cluster more closely

<sup>7981</sup> around the mean of 0.34 than do smaller studies, but large and small

<sup>7982</sup> studies alike have point estimates centered around the mean. That is,

<sup>7983</sup> the funnel plot is symmetric.<sup>11</sup>

<sup>7984</sup> Not all funnel plots are symmetric! figure 16.6 is what happens to our

<sup>7985</sup> hypothetical meta-analysis if all studies with  $p < 0.05$  and positive es-

<sup>7986</sup> timates are published, but only 10% of studies with  $p \geq 0.05$  or with

<sup>7987</sup> negative estimates are published. The introduction of publication bias

<sup>7988</sup> dramatically inflates the pooled estimate from 0.34 to 1.15. Also, there

<sup>7989</sup> appears to be a correlation between studies’ estimates and their stan-

<sup>7990</sup> dard errors, such that smaller studies tend to have larger estimates than

<sup>11</sup> Classic funnel plots look more like figure 16.5). Our version is different in a couple of ways. Most prominently, we don’t have the vertical axis reversed (which we think is confusing). We also don’t have the left boundary highlighted, because we think folks don’t typically select for negative studies.

7991 do larger studies. This correlation is often called **funnel plot asymmetry**  
 7992 because the funnel plot starts to look like a right triangle rather than a  
 7993 funnel. Funnel plot asymmetry *can* be a diagnostic for publication bias,  
 7994 though it isn't always a perfect indicator, as we'll see in the next subsec-  
 7995 tion.

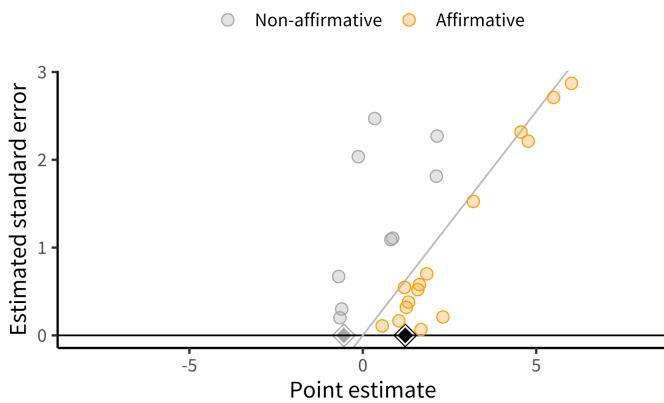


Figure 16.6

A significance funnel plot for the same simulated meta-analysis after publication bias has occurred. Orange points: studies with  $p < 0.05$  and positive estimates. Grey points: studies with  $p \geq 0.05$  or negative estimates. Black diamond: random-effects estimate of  $\hat{\mu}$ .

### 7996 16.2.1 Across-study bias correction

7997 How do we identify and correct bias across studies? Given that some  
 7998 forms of publication bias induce a correlation between studies' point  
 7999 estimates and their standard errors, several popular statistical methods,  
 8000 such as trim-and-fill (Duval and Tweedie 2000) and Egger's regression  
 8001 (Egger et al. 1997) are designed to quantify funnel plot asymmetry.

8002 Funnel plot asymmetry does not always imply that there is publication  
 8003 bias, though. Nor does publication bias always cause funnel plot asym-  
 8004 metry. Sometimes funnel plot asymmetry is driven by genuine differ-

8005 ences in the effects being studied in small and large studies (Egger et al.  
8006 1997; Lau et al. 2006). For example, in a meta-analysis of intervention  
8007 studies, if the most effective interventions are also the most expensive  
8008 or difficult to implement, these highly effective interventions might be  
8009 used primarily in the smallest studies (“small study effects”).

8010 Funnel plots and related methods are best suited to detecting publication  
8011 bias in which (1) small studies with large positive point estimates are  
8012 more likely to be published than small studies with small or negative  
8013 point estimates; and (2) the largest studies are published regardless of  
8014 the magnitude of their point estimates. That model of publication bias  
8015 is sometimes what is happening, but not always!

8016 A more flexible approach for detecting publication bias uses **selection**  
8017 **models**. These models can detect other forms of publication bias that  
8018 funnel plots may not detect, such as publication bias that favors *significant*  
8019 results. We won’t cover these methods in detail here, but we think  
8020 they are a better approach to the question, along with related sensitivity  
8021 analyses.<sup>12</sup>

8022 You may also have heard of “*p*-methods” to detect across-study biases  
8023 such as *p*-curve and *p*-uniform (Simonsohn, Nelson, and Simmons  
8024 2014; van Assen, Aert, and Wicherts 2015). These methods essentially  
8025 assess whether the significant *p*-values “bunch up” just under 0.05,

<sup>12</sup> High-level overviews of selection models are given in McShane, Böckenholt, and Hansen (2016) and Maier, VanderWeele, and Mathur (2022). For more methodological detail, see Hedges (1984), Iyengar and Greenhouse (1988), and Vevea and Hedges (1995). For a tutorial on fitting and interpreting selection models, see Maier, VanderWeele, and Mathur (2022). For sensitivity analyses, see Mathur and VanderWeele (2020b).

8026 which is taken to indicate publication bias. These methods are increas-  
8027 ingly popular in psychology and have their merits. However, they are  
8028 actually simplified versions of selection models (e.g., Hedges 1984) that  
8029 work only under considerably more restrictive settings than the original  
8030 selection models (for example, when there is not heterogeneity across  
8031 studies; McShane, Böckenholdt, and Hansen 2016). For this reason, it is  
8032 usually (although not always) better to use selection models in place of  
8033 the more restrictive *p*-methods.

8034 Going back to our running example, Paluck et al. used a regression-  
8035 based approach to assess and correct for publication bias. This approach  
8036 provided significant evidence of a relationship between the standard  
8037 error and effect size (i.e., an asymmetric funnel plot). Again, this asym-  
8038 metry could reflect publication bias or other sources of correlation be-  
8039 tween studies' estimates and their standard errors. Paluck et al. also used  
8040 this same regression-based approach to try to correct for potential pub-  
8041 lication bias. Results from this model indicated that the bias-corrected  
8042 effect size estimate was close to zero. In other words, even though  
8043 all studies estimated that intergroup contact decreased prejudice, it is  
8044 still possible that there are unpublished studies that did not find this (or  
8045 found that intergroup contact increased prejudice).

 ACCIDENT REPORT

## *Garbage in, garbage out? Meta-analyzing potentially problematic research*

Botox can help eliminate wrinkles. But some researchers have suggested that, when used to paralyze the muscles associated with frowning, Botox may also help treat clinical depression. As surprising as this claim may sound, a quick examination of the literature would lead many to conclude that this treatment works. Studies that randomly assign depressed patients to receive either Botox or saline injections do indeed find that Botox recipients show decreased depression. And when you combine all available evidence in a meta-analysis, you find that this effect is quite large:  $d = 0.83$ , 95% CI [0.52, 1.14].

As Coles et al. (2019) argued though, this estimated effect may be impacted by both within- and between-study bias. First, participants are not supposed to know whether they have been randomly assigned to receive Botox or a control saline injections. But only one of these treatments leads the upper half of your face to be paralyzed! After a couple weeks, you're pretty likely to know whether you received the Botox treatment or control saline injection. Thus, the apparent effect of Botox on depression could instead be a placebo effect.

Second, only 50% of the outcomes that researchers measured were reported in the final publications, raising concerns about selective reporting. Perhaps researchers examining the effects of Botox on depression

only reported the measures that showed a positive effect, not the ones that didn't.

Taken together, these two criticisms suggest that, despite the impressive meta-analytic estimate, the effect of Botox on depression is far from certain.

8047

### 8048 16.3 Chapter summary: Meta-analysis

8049 Taken together, Paluck and colleagues' use of meta-analysis provided  
8050 several important insights that would have been easy to miss in a non-  
8051 quantitative review. First, despite a preponderance of nonsignificant  
8052 findings, intergroup contact interventions were estimated to decrease  
8053 prejudice by on average 0.4 standard deviations. On the other hand,  
8054 there was considerable heterogeneity in intergroup contact effects, sug-  
8055 gesting important moderators of the effectiveness of these interventions.  
8056 And finally, publication bias was a substantial concern, indicating a need  
8057 for follow-up research using a registered report format that will be pub-  
8058 lished regardless of whether the outcome is positive (chapter 11).

8059 Overall, meta-analysis is a key technique for aggregating evidence across  
8060 studies. Meta-analysis allows researchers to move beyond the bias of  
8061 naive techniques like vote counting and toward a more quantitative  
8062 summary of an experimental effect. Unfortunately, a meta-analysis is

8063 only as good as the literature it's based on, so the aspiring meta-analyst

8064 must be aware of both within- and between-study biases!



## DISCUSSION QUESTIONS

1. Imagine that you read the following result in the abstract of a meta-analysis: "In 83 randomized studies of middle school children, replacing one hour of class time with mindfulness meditation significantly improved standardized test scores (standardized mean difference  $\hat{\mu} = 0.05$ ; 95% confidence interval: [0.01, 0.09];  $p < 0.05$ ).". Why is this a problematic way to report on meta-analysis results? Suggest a better sentence to replace this one.
2. As you read the rest of the meta-analysis, you find that the authors conclude that "these findings demonstrate robust benefits of meditation for children, suggesting that test scores improve even when the meditation is introduced as a replacement for normal class time." You recall that the heterogeneity estimate was  $\hat{\tau} = 0.90$ . Do you think that this result regarding the heterogeneity tends to support, or rather tends to undermine, the concluding sentence of the meta-analysis? Why?
3. What kinds of within-study biases would concern you in the meta-analysis described in the prior two questions? How might you assess the credibility of the meta-analyzed studies and of the meta-analysis as a whole in light of these possible biases?
4. Imagine you conduct a meta-analysis on a literature in which statistically significant results in either direction are much more likely to be

published that nonsignificant results. Draw the funnel plot you would expect to see. Is the plot symmetric or asymmetric?

5. Why do you think small studies receive more weight in random-effects meta-analysis than in fixed-effects meta-analysis? Can you see why this is true mathematically based on the equations given above, and can you also explain the intuition in simple language?

8066

## READINGS

- A nice, free textbook with lots of good code examples: Harrer, Mathias, Pim Cuijpers, Furukawa Toshi A, and David D. Ebert (2021). *Doing Meta-Analysis with R: A Hands-On Guide*. Chapman & Hall/CRC Press. Available free online at [https://bookdown.org/MathiasHarrer/Doing\\_Meta\\_Analysis\\_in\\_R](https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R).

8067

## References

- Allport, Gordon Willard, Kenneth Clark, and Thomas Pettigrew. 1954. *The Nature of Prejudice*. Addison-Wesley.
- Boisjoly, Johanne, Greg J. Duncan, Michael Kremer, Dan M. Levy, and Jacqueline Eccles. 2006. “Empathy or Antipathy? The Impact of Diversity.” *American Economic Review* 96 (5): 1890–1905.
- Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to Meta-Analysis*. John Wiley & Sons.
- Braginsky, Mika, Maya Mathur, and Tyler J. VanderWeele. 2023. *Publication Bias in Meta-Analysis*. Oxford University Press.

8069

- tionBias: Sensitivity Analysis for Publication Bias in Meta-Analyses.* <https://CRAN.R-project.org/package=PublicationBias>.
- Brockwell, Sarah E., and Ian R. Gordon. 2001. “A Comparison of Statistical Methods for Meta-Analysis.” *Statistics in Medicine* 20 (6): 825–840.
- Clunies-Ross, Graham, and Kris O’Meara. 1989. “Changing the Attitudes of Students towards Peers with Disabilities.” *Australian Psychologist* 24 (2): 273–284.
- Coles, Nicholas A., Jeff T. Larsen, Joyce Kurabayashi, and Ashley Kuelz. 2019. “Does Blocking Facial Feedback via Botulinum Toxin Injections Decrease Depression? A Critical Review and Meta-Analysis.” *Emotion Review* 11 (4): 294–309.
- Cooper, Harris, and Erika A. Patall. 2009. “The Relative Benefits of Meta-Analysis Conducted with Individual Participant Data Versus Aggregated Data.” *Psychological Methods* 14 (2): 165–176. <https://doi.org/10.1037/a0015565>.
- DerSimonian, Rebecca, and Nan Laird. 1986. “Meta-Analysis in Clinical Trials.” *Controlled Clinical Trials* 7 (3): 177–188.
- Duval, Sue, and Richard Tweedie. 2000. “Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis.” *Biometrics* 56 (2): 455–463. <https://doi.org/10.1111/j.0006-341X.2000.00455.x>.
- Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. “Bias in Meta-Analysis Detected by a Simple, Graphical Test.” *British Medical Journal* 315 (7109): 629–634. <https://doi.org/10.1136/bmj.315.7109.629>.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. “Publication Bias in the Social Sciences: Unlocking the File Drawer.” *Science* 345 (6203): 1502–1505. <https://doi.org/10.1126/science.1255484>.

Goh, Jin X., Judith A. Hall, and Robert Rosenthal. 2016. “Mini Meta-Analysis of Your Own Studies: Some Arguments on Why and a Primer on How.” *Social and Personality Psychology Compass* 10 (10): 535–549.

Goldstein, Noah J., Robert B. Cialdini, and Vladas Griskevicius. 2008. “A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels.” *Journal of Consumer Research* 35 (3): 472–482.

Grant, Maria J., and Andrew Booth. 2009. “A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies.” *Health Information & Libraries Journal* 26 (2): 91–108.

Harrer, Mathias, Pim Cuijpers, Furukawa Toshi A, and David D. Ebert. 2021. *Doing Meta-Analysis with R: A Hands-on Guide*. Chapman & Hall/CRC Press.

Hedges, Larry V. 1984. “Estimation of Effect Size under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences.” *Journal of Educational Statistics* 9 (1): 61–85. <https://doi.org/10.3102/10769986009001061>.

Hedges, Larry V., Elizabeth Tipton, and Matthew C. Johnson. 2010. “Robust Variance Estimation in Meta-Regression with Dependent Effect Size Estimates.” *Research Synthesis Methods* 1 (1): 39–65.

Iyengar, Satish, and Joel B. Greenhouse. 1988. “Selection Models and the File Drawer Problem.” *Statistical Science*, 109–117.

Lau, Joseph, John P. A. Ioannidis, Norma Terrin, Christopher H. Schmid, and

Ingram Olkin. 2006. “The Case of the Misleading Funnel Plot.” *British Medical Journal* 333 (7568): 597–600. <https://doi.org/10.1136/bmj.333.7568.597>.

Lefebvre, Carol, Julie Glanville, Simon Briscoe, Anne Littlewood, Chris Marshall, Maria-Inti Metzendorf, Anna Noel-Storr, et al. 2019. “Searching for and Selecting Studies.” In *Cochrane Handbook for Systematic Reviews of Interventions*, edited by Julian P T Higgins, J. Thomas, M. Chandler, T. Cumpston, M. J. Page Li, and V. A. Welch, 67–107. Wiley-Blackwell. <https://doi.org/10.1002/9781119536604.ch4>.

Maier, Maximilian, Tyler J. VanderWeele, and Maya B. Mathur. 2022. “Using Selection Models to Assess Sensitivity to Publication Bias: A Tutorial and Call for More Routine Use.” *Campbell Systematic Reviews* 18 (3): e1256.

Mathur, Maya B., and Tyler J. VanderWeele. 2019. “New Metrics for Meta-Analyses of Heterogeneous Effects.” *Statistics in Medicine* 38 (8): 1336–1342.

Mathur, Maya B., and Tyler J. VanderWeele. 2020a. “Robust Metrics and Sensitivity Analyses for Meta-Analyses of Heterogeneous Effects.” *Epidemiology* 31 (3): 356–358.

Mathur, Maya B., and Tyler J. VanderWeele. 2020b. “Sensitivity Analysis for Publication Bias in Meta-Analyses.” *Journal of the Royal Statistical Society: Series C* 5 (69): 1091–1119.

Mathur, Maya B., and Tyler J. VanderWeele. 2021. “Estimating Publication Bias in Meta-Analyses of Peer-Reviewed Studies: A Meta-Meta-Analysis across Disciplines and Journal Tiers.” *Research Synthesis Methods* 12 (2): 176–191.

Mathur, Maya B., and Tyler J. VanderWeele. 2022. “Methods to Address Confounding and Other Biases in Meta-Analyses: Review and Recommendations.” *Annual Review of Public Health* 1 (43).

McShane, Blakeley B., Ulf Böckenholt, and Karsten T. Hansen. 2016. “Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes.” *Perspectives on Psychological Science* 11 (5): 730–749. <https://doi.org/10.1177/1745691616662243>.

McShane, Blakeley B., and David Gal. 2017. “Statistical Significance and the Dichotomization of Evidence.” *Journal of the American Statistical Association* 112 (519): 885–895. <https://doi.org/10.1080/01621459.2017.1289846>.

Nelson, Nanette, Robert Rosenthal, and Ralph L. Rosnow. 1986. “Interpretation of Significance Levels and Effect Sizes by Psychological Researchers.” *American Psychologist* 41 (11): 1299–1301. <https://doi.org/10.1037/0003-066X.41.11.1299>.

Paluck, Elizabeth Levy, Seth A. Green, and Donald P. Green. 2019. “The Contact Hypothesis Re-Evaluated.” *Behavioural Public Policy* 3 (2): 129–158.

Pustejovsky, James E., and Elizabeth Tipton. 2021. “Meta-Analysis with Robust Variance Estimation: Expanding the Range of Working Models.” *Prevention Science* 23 (2022): 425–438. <https://doi.org/10.1007/s11121-021-01246-3>.

Riley, Richard D., Julian P. T. Higgins, and Jonathan J. Deeks. 2011. “Interpretation of Random Effects Meta-Analyses.” *British Medical Journal* 342.

Scheibehenne, Benjamin, Tahira Jamil, and Eric-Jan Wagenmakers. 2016. “Bayesian Evidence Synthesis Can Reconcile Seemingly Inconsistent Re-

sults: The Case of Hotel Towel Reuse.” *Psychological Science* 27 (7): 1043–1046.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. “P-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143 (2): 534–547.

Sterne, Jonathan A. C., Miguel A. Hernán, Barnaby C. Reeves, Jelena Savović, Nancy D. Berkman, Meera Viswanathan, David Henry, et al. 2016. “ROBINS-I: A Tool for Assessing Risk of Bias in Non-Randomised Studies of Interventions.” *British Medical Journal* 355. <https://doi.org/doi.org/10.1136/bmj.i4919>.

Thompson, Simon G., and Julian P. T. Higgins. 2002. “How Should Meta-Regression Analyses Be Undertaken and Interpreted?” *Statistics in Medicine* 21 (11): 1559–1573.

Tipton, Elizabeth. 2015. “Small Sample Adjustments for Robust Variance Estimation with Meta-Regression.” *Psychological Methods* 20 (3): 375–393. <https://doi.org/10.1037/met0000011>.

Tsuji, Sho, Alejandrina Cristia, Michael C. Frank, and Christina Bergmann. 2020. “Addressing Publication Bias in Meta-Analysis.” *Zeitschrift für Psychologie* 228 (1): 50–61. <https://doi.org/10.1027/2151-2604/a000393>.

van Assen, Marcel A. L. M., Robbie van Aert, and Jelte M. Wicherts. 2015. “Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies.” *Psychological Methods* 20 (3): 293–309.

Vevea, Jack L., and Larry V. Hedges. 1995. “A General Linear Model for Estimating Effect Size in the Presence of Publication Bias.” *Psychometrika* 60 (3): 419–435.

Viechtbauer, Wolfgang. 2010. “Conducting Meta-Analyses in R with the metafor Package.” *Journal of Statistical Software* 36 (3): 1–48. <https://doi.org/10.18637/jss.v036.i03>.

Wang, Chia-Chun, and Wen-Chung Lee. 2019. “A Simple Method to Estimate Prediction Intervals and Predictive Distributions: Summarizing Meta-Analyses beyond Means and Confidence Intervals.” *Research Synthesis Methods* 10 (2): 255–266.

## 8076 17 CONCLUSION

8077 You've made it to the end of *Experimentology*, our (sometimes opinion-  
8078 ated) guide to how to run good psychology experiments. In this book  
8079 we've tried to present a unified approach to the why and how of run-  
8080 ning experiments. This approach begins with the goal of doing experi-  
8081 ments:

8082 Experiments are intended to make maximally unbiased,  
8083 generalizable, and precise estimates of specific causal  
8084 effects.

8085 This formulation isn't exactly how experiments are talked about in the  
8086 broader field, but we hope you've started to see some of the rationale  
8087 behind this approach. In this final chapter, we will briefly discuss  
8088 some aspects of our approach, as well how this approach connects  
8089 with our four themes, TRANSPARENCY, MEASUREMENT PRECISION, BIAS

8090 REDUCTION, and GENERALIZABILITY. We'll end by mentioning some  
8091 exciting new trends in the field that give us hope about the future of  
8092 experimentology and psychology more broadly.

### 8093 17.1 *Summarizing our approach*

8094 The *Experimentology* approach is grounded in both an appreciation of  
8095 the power of experiments to reveal important aspects about human psy-  
8096 chology and also an understanding of the many ways that experiments  
8097 can fail. In particular, the “replication crisis” (chapter 3) has revealed  
8098 that small samples, a focus on dichotomous statistical inference, and a  
8099 lack of transparency around data analysis can lead to a literature that is  
8100 often neither reproducible nor replicable. Our approach is designed to  
8101 avoid these pitfalls.

8102 We focus on MEASUREMENT PRECISION in service of measuring causal ef-  
8103 fects. The emphasis on causal effects stems from an acknowledgement  
8104 of the key role of experiments in establishing causal inferences (chap-  
8105 ter 1) and the importance of causal relationships to theories (chapter 2).

8106 In our statistical approach, we focus on estimation (chapter 5) and mod-  
8107 eling (chapter 7), helping us to avoid some of the fallacies that come  
8108 along with dichotomous inference (chapter 6). We choose measures to  
8109 maximize reliability (chapter 8). We prefer simple, within-participant

8110 experimental designs because they typically result in more precise es-  
8111 timates (chapter 9). And we think meta-analytically about the over-  
8112 all evidence for a particular effect beyond our individual experiment  
8113 (chapter 16).

8114 Further, we recognize the presence of many potential sources of bias in  
8115 our estimates, leading us to focus on BIAS REDUCTION. In our measure-  
8116 ments, we identify arguments for the validity of our measures, decreas-  
8117 ing bias in estimation of the key constructs of interest (chapter 8); in our  
8118 designs we seek to minimize bias due to confounding or experimenter  
8119 effects (chapter 9). We also try to minimize the possibility of bias in  
8120 our decisions about data collection (chapter 12) and data analysis (chap-  
8121 ter 11). Finally, we recognize the possibility of bias in literatures as a  
8122 whole and consider ways to compensate in our estimates (chapter 16).

8123 Finally, we consider GENERALIZABILITY throughout the process. We the-  
8124 orize with respect to a particular population (chapter 2) and select our  
8125 sample in order to maximize the generalizability of our findings to that  
8126 target population (chapter 10). In our statistical analysis, we take into  
8127 account multiple dimensions of generalizability, including across par-  
8128 ticipants and experimental stimulus items (chapter 7). And in our re-  
8129 porting, we contextualize our findings with respect to limits on their  
8130 generalizability (chapter 14).

8131 Woven throughout this narrative is the hope that embracing TRANS-  
8132 PARENCEY throughout the experimental process will help you maximize  
8133 your work. Not only is sharing your work openly an ethical respon-  
8134 sibility (chapter 4), but it's also a great way to minimize errors while  
8135 creating valuable products that both advance scientific progress and ac-  
8136 celerate your own career (chapter 13).

## 8137 *17.2 Forward the field*

8138 We have focused especially on reproducibility and replicability issues,  
8139 but we've learned at various points in this book that there's a replica-  
8140 tion crisis (Open Science Collaboration 2015), a theory crisis (Oberauer  
8141 and Lewandowsky 2019), and a generalizability crisis (Yarkoni 2020) in  
8142 psychology. Based on all these crises, you might think that we are pes-  
8143 simistic about the future of psychology. Not so.

8144 There have been tremendous changes in psychological methods since  
8145 we started teaching Experimental Methods in 2012. When we began,  
8146 it was common for incoming graduate students to describe the ram-  
8147 pant *p*-hacking they had been encouraged to do in their undergraduate  
8148 labs. Now, students join the class aware of new practices like preregis-  
8149 tration and cognizant of problems of generalizability and theory build-  
8150 ing. It takes a long time for a field to change, but we have seen tremen-

8151 dous progress at every level—from government policies requiring trans-  
8152 parency in the sciences all the way down to individual researchers’ adop-  
8153 tion of tools and practices that increase the efficiency of their work and  
8154 decrease the chances of error.

8155 One of the most exciting trends has been the rise of metascience, in  
8156 which researchers use the tools of science to understand how to make  
8157 science better (Tom E. Hardwicke et al. 2020). Reproducibility and  
8158 replicability projects (reviewed in chapter 3) can help us measure the  
8159 robustness of the scientific literature. In addition, studies that evalua-  
8160 ate the impacts of new policies (e.g., Tom E. Hardwicke et al. 2018)—  
8161 can help stakeholders like journal editors and funders make informed  
8162 choices about how to push the field toward more robust science.

8163 In addition to changes that correct methodological issues, the last ten  
8164 years have seen the rise of “big team science” efforts that advance the  
8165 field in new ways (Coles et al. 2022). Collaborations such as the Psy-  
8166 chological Science Accelerator (Moshontz et al. 2018) and ManyBabies  
8167 (Frank et al. 2017) allow hundreds of researchers from around the world  
8168 to come together to run shared projects. These projects are enabled by  
8169 open science practices like data and code sharing, and they provide a  
8170 way for researchers to learn best practices via participating. In addition,  
8171 by including broader and more diverse samples, they can help address

8172 challenges around generalizability (Klein et al. 2018).

8173 Finally, the last ten years have seen huge progress in the use of statisti-  
8174 cal models both for understanding data (McElreath 2018) and for de-  
8175 scribing specific psychological mechanisms (Ma, Körding, and Goldre-  
8176 ich 2022). In our own work, we have used these models extensively and  
8177 we believe that they provide an exciting toolkit for building quantita-  
8178 tive theories that allow us to explain and to predict the human mind.

### 8179 *17.3 Final thoughts*

8180 Doing experiments is a craft, one that requires practice and attention.  
8181 The first experiment you run will have limitations and issues. So will  
8182 the 100th. But as you refine your skills, the quality of the studies you  
8183 design will get better. Further, your own ability to judge others' exper-  
8184 iments will improve as well, making you a more discerning consumer  
8185 of empirical results. We hope you enjoy this journey!

**8186 References**

- Coles, Nicholas A., J. Kiley Hamlin, Lauren L. Sullivan, Timothy H. Parker, and Drew Altschul. 2022. "Build Up Big-Team Science." *Nature Publishing Group*.
- Frank, Michael C., Elika Bergelson, Christina Bergmann, Alejandrina Cristia, Caroline Floccia, Judit Gervain, J. Kiley Hamlin, et al. 2017. "A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building." *Infancy* 22 (4): 421–435.
- Hardwicke, Tom E., Maya B. Mathur, Kyle Earl MacDonald, Gustav Nilsonne, George Christopher Banks, Mallory Kidwell, Alicia Hofelich Mohr, et al. 2018. "Data Availability, Reusability, and Analytic Reproducibility: Evaluating the Impact of a Mandatory Open Data Policy at the Journal *Cognition*." *Royal Society Open Science* 5. <https://doi.org/10.1098/rsos.180448>.
- Hardwicke, Tom E., Stylianos Serghiou, Perrine Janiaud, Valentin Danchev, Sophia Crüwell, Steven N. Goodman, and John P. A. Ioannidis. 2020. "Calibrating the Scientific Ecosystem through Meta-Research." *Annual Review of Statistics and Its Application* 7 (1): 11–37. <https://doi.org/10.1146/annurev-statistics-031219-041104>.
- Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams Jr, Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating Variation in Replicability across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–490.
- Ma, Wei Ji, K. Körding, and Daniel Goldreich. 2022. *Bayesian Models of Perception and Action: An Introduction*. unpublished.
- McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*.

- ples in R and Stan.* Chapman and Hall/CRC.
- Moshontz, Hannah, Lorne Campbell, Charles R. Ebersole, Hans IJzerman, Heather L. Urry, Patrick S. Forscher, Jon E. Grahe, et al. 2018. “The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network.” *Advances in Methods and Practices in Psychological Science* 1 (4): 501–515.
- Oberauer, Klaus, and Stephan Lewandowsky. 2019. “Addressing the Theory Crisis in Psychology.” *Psychonomic Bulletin & Review* 26 (5): 1596–1618.
- Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251).
- Yarkoni, Tal. 2020. “The Generalizability Crisis.” *Behavioral and Brain Sciences* 45:1–37.

## 8189 A INSTRUCTOR'S GUIDE

### 8190 *A.1 Introduction*

8191 This is an instructor's guide to conducting replication projects in courses.  
8192 In addition to benefiting the field in ways that have been previously  
8193 discussed by some of the authors of this book (e.g., Hawkins et al. 2018;  
8194 Frank and Saxe 2012), replication-based courses can additionally benefit  
8195 students in these courses. In this guide, we will describe these benefits,  
8196 explore different ways in which courses may be modified depending on  
8197 student level and resources, and provide some guidelines and examples  
8198 to help you set up the logistics of your course.

### 8199 *A.2 Why Teach a Project-Based Course?*

8200 Over the years, we have observed many ways in which our replication-  
8201 based courses benefited students above and beyond a more traditional

8202 lecture and problem set-based course. Some of these benefits include:

- 8203 – **Student interest:** Since each student will be free to replicate a  
study that is aligned with their research interests, this freedom fa-  
cilitates a more direct application of course methods and lessons  
8205 to a project that is interesting to each student.
- 8206 – **Usefulness:** If this course is taught in the first year of the program  
8207 (as recommended), students may use their replication project as  
8208 a way to establish robustness of a phenomenon before building  
8209 studies on top of it.
- 8210 – **Realism:** Practice datasets that are typically provided for course  
8211 exercises lack the complexity and messiness of real data. By con-  
8212 ducting a replication project and dealing with real data, students  
8213 learn to apply the tools provided in the course in a way that more  
8214 closely demonstrates their usefulness beyond the course.
- 8215 – **Intuition:** Presentations of replication outcomes across the class  
8216 along with a discussion of what factors seemed to predict these  
8217 outcomes helps students develop a better intuition when reading  
8218 the literature for how likely studies are to replicate.
- 8219 – **Perspective:** Frustrating experiences with ambiguity (whether re-  
8220 garding experimental methods, materials, or analyses) can moti-  
8221 vate students to adopt best practices for their own future studies.

8223 A project-based course may look very different depending on student  
8224 level (undergraduate vs graduate/post-doc level) and availability of re-  
8225 sources at your institution for a course like this, namely in terms of TA  
8226 support and course funding (for data collection). For most of this guide,  
8227 we will assume that you have a similar setup to ours (i.e., teaching at the  
8228 graduate/post-doc level and have course funding and TAs to support  
8229 the course), but we have also spent some time considering ways to ad-  
8230 just the course to fit different student levels and availability of resources  
8231 (see “Scenarios for different course layouts”).

### 8232 *A.3 Logistics*

#### 8233 *A.3.1 Syllabus considerations*

8234 If it is your first time teaching this course, you may want to decide ahead  
8235 of time whether your course will mainly focus on content, or whether  
8236 you will cover *both* content and relevant practical skills. For instance,  
8237 if the course is for undergraduate students, you may decide to focus  
8238 mainly on content, whereas if the course is for graduate students, they  
8239 may find it more useful if the course covers both content and practical  
8240 skills they can use in their research.

8241 Another important consideration is how long your course will be. De-  
8242 pending on whether your university operates on quarters or semesters,  
8243 the pace of the course will differ. For Psych 251, since we are on the  
8244 quarter system, we use the 10-week schedule shown below. However,  
8245 we have also adapted this schedule to a 16-week course given that it  
8246 better represents a majority of other institutions' academic calendars. At  
8247 the end of this chapter, we give a set of sample class schedules.

### 8248 *A.3.2 Grading*

8249 Depending on your course format and teaching philosophy, you may  
8250 have preferred grading criteria. As a point of reference, in Psych 251,  
8251 we wanted to encompass both the assignments (problemsets and project  
8252 components) as well as actual course attendance and participation. In  
8253 addition, because the replication project is a central part of the course,  
8254 we weighted the project components slightly more than the problem  
8255 sets:

- 8256 – 40%: Problem sets (four, at 10% each)
- 8257 – 50%: Final project components, including presentations, data col-  
8258 lection, analysis, and writeup
- 8259 – 10%: Attendance and participation in class

8260 *A.3.3 Course budget*

8261 For our course, we usually receive around US\$1,000 for course  
8262 funding from the Psychology Department. In addition, when students  
8263 from other departments are enrolled, we have been lucky to receive  
8264 additional funding from those departments as well, to further support  
8265 the course. Still, making sure that the course funds cover all students'  
8266 projects is one of the most challenging parts of the course. Assuming  
8267 you have a budget to work with, here are some lessons we've learned  
8268 along the way regarding budgeting (and if you don't have any funding,  
8269 please refer to the section titled "Course Funding" under "Scenarios  
8270 for different course layouts"):

- 8271 – Before students pick their study to replicate, provide them with  
8272 an estimate of how many participant hours they will be able to  
8273 receive for their project
- 8274 – As soon as students pick a study for their replication project, help  
8275 each student run a power analysis to confirm that replicating the  
8276 study would be within the budget (TAs can help with this)
- 8277 – If a student feels strongly about a study that does not fit within the  
8278 budget, consider the following ways to adjust the study: (1) Can  
8279 the study be made shorter by cutting out unnecessary measures?  
8280 (2) If it is a multi-trial study, can the number of trials be reduced?

8281 (3) Would their advisors be willing to provide additional funding?

8282 (4) can the study be run on university participant pools?

8283 – As mentioned above, if there are students from other departments

8284 who are enrolled in your course, one possibility to obtain more

8285 funding is to reach out to the heads of those departments to see

8286 whether they would be willing to help support your course.

8287 Once all projects have been approved as within-budget, we encourage

8288 you to create a shared spreadsheet containing each student's name, so

8289 that they can fill in the details of their replication project. Ultimately,

8290 this will help ensure that students are paying fair wages to their parti-

8291 pants and keep track of how the course funds are being divided up.

8292 *A.3.4 Course-related Institutional Review Board application*

8293 While it may be possible to apply for individual IRB approval for each

8294 student's project, we recommend applying for course-wide standard

8295 IRB approval for all replication projects that are conducted in your class.

8296 Contacting your review board early in the planning stages of the course

8297 should clarify what options you have available.

8298 One important thing to remember when students run their individual

8299 projects is that they should have the course-wide consent form at the

8300 beginning of their studies (and TAs should check this when they review  
8301 the paradigms). For reference, this is the consent form that each student  
8302 is required to post at the beginning of their study:

8303 “By answering the following questions, you are participating in a study  
8304 being performed by cognitive scientists in the Stanford Department of  
8305 Psychology. If you have questions about this research, please contact us  
8306 at stanfordpsych251@gmail.com. You must be at least 18 years old to  
8307 participate. Your participation in this research is voluntary. You may  
8308 decline to answer any or all of the following questions. You may de-  
8309 cline further participation, at any time, without adverse consequences.  
8310 Your anonymity is assured; the researchers who have requested your  
8311 participation will not receive any personal information about you.”

8312 *A.4 Scenarios for different course layouts*

8313 Now that we have covered the standard format of the course, we want to  
8314 now turn our attention to ways in which this format can be tweaked in  
8315 order to fit different needs and resources. We have organized this section  
8316 into two main categories: student level and course resources (such as  
8317 TAs and course funding).

*A.4.1 Student level*

8318 While Psych 251 at Stanford is geared toward graduate students (and  
8319 is currently a required class for entering first-year graduate students in  
8320 the Psychology Department), we also accept advanced undergraduate  
8321 students as well as graduate students from other departments (e.g., ed-  
8322 ucation, human-computer interaction, philosophy, computer science).  
8323

8324 On the first day of our course, we tell students that they should be com-  
8325 fortable with two of the three following topics:

- 8326 1. Some knowledge of psychological experimentation & subject  
8327 matter
- 8328 2. Statistical programming: things like functions and variables
- 8329 3. Basic statistics like ANOVA and t-test

8330 If students are only comfortable with one of the three topics above, we  
8331 warn them ahead of time that the course may demand more time from  
8332 them than the average student.

8333 Now, if you are planning on catering this course for undergraduate stu-  
8334 dents, chances are that they have had less exposure to these topics over-  
8335 all, so there are multiple ways to calibrate the course accordingly:

8336 **Prerequisites:** Require students to have completed courses that cover at  
8337 least two of the three topics mentioned above (i.e., a psychology class, a  
8338 class that covers statistical programming, a class that covers basic statis-  
8339 tics, any two of the three).

8340 **Pace:** Unlike Psych 251, where the entire course only lasts 10 weeks,  
8341 a class for undergraduates may benefit from a slower pace, allowing  
8342 more time to cover the foundational principles before diving into the  
8343 project. For instance, the course could be held over multiple academic  
8344 semesters/quarters, with the project goal of Course #1 being choosing  
8345 and planning the replication study, and the project goal of Course #2  
8346 being the execution and interpretation of the replication.

8347 **Pair-group-based projects:** In our course, each student is required to  
8348 conduct their own replication project. However, this structure may  
8349 be overwhelming for undergraduate students who may have less con-  
8350 fidence taking on an entire replication project by themselves. One op-  
8351 tion that may alleviate this pressure is to have students conduct these  
8352 projects as pairs or as small teams, so that they can collectively draw on  
8353 each others' strengths. When assigning these pairs or teams, it may be  
8354 especially helpful to try to ensure a relatively even balance of students  
8355 who are confident in each of the three areas outlined above (psychology,  
8356 statistical programming, basic statistics).

8357 Now that we've offered a few suggestions to address different student  
8358 levels, let's dive into the issue of course resources.

8359 *A.4.2 Course resources*

8360 We think there are two main ways in which your course may have differ-  
8361 ent resources from our model: in terms of course assistance (i.e., teach-  
8362 ing assistants), and in terms of course funding for student projects. We'll  
8363 explore ways to work around each of these in this section.

8364 **Teaching assistants.** As a point of comparison, in general, two to three  
8365 teaching assistants are allocated to Psych 251, which enrolls about 36 stu-  
8366 dents, which comes out to about 12–18 students per TA. Since a project-  
8367 based course requires individual attention and feedback, we would rec-  
8368 ommend against a student-TA ratio that is much higher than that. That  
8369 means that if you know you will have just one TA for the class, you  
8370 should think about reducing the enrollment cap accordingly. But what  
8371 if you have *no* TAs? With some adjustments, there are still ways you can  
8372 make the course work sans-TA; we outline a few ideas below:

- 8373 1. **Peer grading.** As an instructor with no TAs, the area that will  
8374 require the biggest lift in terms of time and attention is grading.  
8375 One way to overcome this is to introduce a peer-grading system,

8376 in which students grade each others' work. If you choose this  
8377 route, two things that may encourage fair grading among your  
8378 students is to (1) distribute a clear and specific rubric that reduces  
8379 the amount of subjectivity in the grading process as much as pos-  
8380 sible, and (2) anonymize the assignments so that students do not  
8381 know whose assignment they are grading. If possible, it may again  
8382 be beneficial to assign grading pairs that consist of students that  
8383 are relatively knowledgeable in different areas, so that they can  
8384 provide feedback that address weak points in each others' work.

8385 **2. Collective troubleshooting.** The second most time intensive area  
8386 you will have to make up for is the amount of troubleshooting  
8387 you may have to do for students who run into issues implement  
8388 their projects, anywhere from getting GitHub and R Markdown  
8389 up and running on their devices, to trouble with data collection  
8390 on Mechanical Turk. One way to encourage communal support  
8391 among your students is to set up a central discussion board for the  
8392 course (e.g., Piazza or a course channel on Slack) where students  
8393 can publicly (but anonymously, if desired) post issues they are run-  
8394 ning into. Then, you can offer extra credit to students who help  
8395 troubleshoot these issues, in order to further incentivize collec-  
8396 tive troubleshooting. There will likely still be issues that cannot  
8397 be addressed by the students, but this system at least frees up your

8398 time to focus your attention on those that only *you* can address.

8399 3. **Single class-wide project.** Finally, if the collective grading  
8400 and troubleshooting methods outlined above do not cut down  
8401 on enough time, you could consider walking through a single  
8402 replication project as a class.<sup>1</sup> To make a single-project course  
8403 work, you could have students nominate studies they would  
8404 like to replicate as a class, and then have them vote on the final  
8405 choice. Once the target study has been selected, every student  
8406 can individually carry out all the steps of the project, including  
8407 preregistering and writing up the analysis script. Then, setting  
8408 up and running the data collection phase can happen during  
8409 class, and once data has been collected, you can distribute it to  
8410 the students for them to run it through their analysis script and  
8411 interpret the result. Whether you choose to have students grade  
8412 each others' work or whether you grade their work yourself, the  
8413 fact that the project is standardized should cut down on a lot of  
8414 the time you would otherwise spend learning about the details  
8415 of every individual project.

8416 **Course funding.** In addition to availability of TAs, another way in  
8417 which your course may be different from ours is in terms of course  
8418 funding. If you have little or not funding for your course (even after

<sup>1</sup> This approach does cancel out some of the benefits of a project-based course we mentioned at the start—namely, the project will likely no longer fit each student's specific research interest, so there may be less benefit in terms of specific student interest and usefulness for their program of research, but the other two benefits of realism and intuition (especially if the project is discussed in the context of other replication findings) still stand.

8419 reaching out to relevant members of your department or institution),

8420 we suggest the following adjustments:

8421 1. **Pair-group-based projects.** Similarly to suggestion 3 for address-

8422 ing different student levels, one option for limited course bud-

8423 gets is to have students conduct the replication projects as pairs or

8424 teams to reduce the cost of data collection. This structure may

8425 have the added benefit of encouraging students to problem-solve

8426 together. Alternatively, each student in the pairs or teams could

8427 complete each step of the replication individually (e.g., writing

8428 up the report, analyzing the data, interpreting the result), which

8429 would ensure that each student takes full responsibility for every

8430 step of the project. This structure may also provide opportunities

8431 for interesting discussions at the end of the course around ana-

8432 lytic reproducibility, especially if students in the same teams (with

8433 the same dataset) differed in the conclusions they drew about the

8434 replication outcome.

8435 2. **Funding from advisors.** In some cases, students come to us with

8436 target studies that require more funding than we are able to al-

8437 locate, but that they feel particularly invested in (e.g., because of

8438 how relevant the study is to their line of research). Once we rule

8439 out other ways of making the study fit our budget (e.g., dropping

extra control conditions, running a subset of the study), we often ask students whether their advisor would be willing to fund the study. We have found that advisors are often willing to do this, especially if the replication could serve an important role in the development of the student's research program. Similarly, one way to reduce the burden on a limited course budget would be to encourage all students to first ask their advisors about whether they would be willing to fund part or all of the data collection for the replication. While chances are that some advisors will be unwilling or unable to do this, there should still be a meaningful reduction in the number of projects the course will need to fund.

3. **Reproduce a replication.** The suggestions above apply if you at least have *some* amount of course funding, but what if you have *no* funding at all? While there are obvious limitations to this solution, one suggestion is to have students reproduce past public replications. For instance, our course GitHub page<sup>2</sup> contains public repositories of all past replication projects that have been conducted in our course. Since the data for each replication project is available in these repositories, you could provide each of your students with a dataset and the original paper associated with it, and assign them to reproduce the results of the replication. Students should then be able to follow each step of the replication

<sup>2</sup> <https://github.com/psych251>

8462 project described below (e.g., writing the report, identifying the  
8463 key analysis, running the analysis). This format will only work if  
8464 students do not view the original final replication reports that are  
8465 posted publicly for their project, so it may be necessary to be clear  
8466 about this at the beginning of the course.

8467 For those of you who are working with a different course format  
8468 (whether in terms of student level or course resources), we hope these  
8469 suggestions were useful. If you try out a new idea in your course that  
8470 you found helpful, we would be thrilled if you shared them with us!

#### 8471 *A.5 Sample course schedules*

8472 The sample syllabi laid out below are categorized along the following  
8473 decisions: (1) material: whether the course focuses on just content or  
8474 both content and skills, and (2) duration: whether the course is 10 weeks  
8475 long or 16 weeks long.

8476 For undergraduate instructors, we have labeled advanced topics in pur-  
8477 ple. We expect that these topics are best suited for advanced under-  
8478 graduate students. As for content around statistics (e.g., estimation, in-  
8479 ference), instructors should decide how much of this content to teach,  
8480 depending on how prepared students have been in previous classes.

8481 A.5.1 10 weeks

Table A.1  
A sample 10-week syllabus with both skills and content materials.

Week	Day	Topic	Chapter	Appendix
1	M	Class introduction	1	
1	W	Theories	2	
1	F	Version control		B
2	M	Reproducible reports	14	C
2	W	Tidyverse tutorial		D
2	F	Tidyverse tutorial continued (with TAs)		
3	M	Measurement, reliability, and validity	8	
3	W	Design of experiments	9	
3	F	Sampling	10	
4	M	Project management	13	
4	W	Experiments 1: Simple survey experiments using Qualtrics		
4	F	Experiments 2: Project-specific implementation (TAs)		
5	M	Estimation	5	
5	W	Inference	6	
5	F	Sample size planning		
6	M	Survey design		
6	W	Midterm presentations 1		
6	F	Midterm presentations 2		

7	M	Preregistration	11
7	W	Meta-analysis	16
7	F	Open science	3
8	M	Visualization 1	15 E
8	W	Visualization 2	
8	F	Exploratory data analysis workshop	
9	M	Sampling, representativeness, and generalizability	4
9	W	Data and participant ethics	12
9	F	Authorship and research ethics	
10	M	Open discussion	17
10	W	Final Project presentations 1	
10	F	Final Project presentations 2	

## 8482 A.5.2 10 weeks, content only

Table A.2  
A sample 10-week syllabus with only content materials.

Week	Day	Topic	Chapter
1	M	Class introduction	1
1	W	Theories	2
1	F	Replication and reproducibility	3
2	M	Open Science	
2	W	Measurement	8
2	F	Design of experiments 1	9
3	M	Design of experiments 2	
3	W	Sampling	10
3	F	Experimental strategy	
4	M	Preregistration	11
4	W	Data collection	12
4	F	Visualization 1	15
5	M	Visualization 2	
5	W	Midterm exam	
5	F	Introduction to statistics	
6	M	Estimation 1	5
6	W	Estimation 2	
6	F	Inference 1	6

---

7	M	Inference 2	
7	W	Models 1	7
7	F	Models 2	
8	M	Meta-analysis	16
8	W	Project management	13
8	F	[Instructor-specific topics]	
9	M	Sampling, representativeness, and generalizability	4
9	W	Data and participant ethics	12
9	F	Authorship and research ethics	
10	M	Conclusion	17
10	W	Conclusion	
10	F	Final exam	

---

## 8483 A.5.3 16 weeks

Table A.3  
A sample 16-week syllabus with both skills and content materials.

Week	Day	Topic	Chapter	Appendix
1	1	Class introduction	1	
1	2	Theories	2	
2	1	Version control		B
2	2	Reproducible reports	14	C
3	1	Tidyverse tutorial		D
3	2	Tidyverse tutorial continued (with TAs)		
4	1	Measurement, reliability, and validity	8	
4	2	Design of experiments	9	
5	1	Sampling	10	
5	2	Project management	13	
6	1	Experiments 1: Simple survey experiments using Qualtrics		
6	2	Experiments 2: Project-specific implementation (TAs)		
7	1	Estimation	5	
7	2	Inference	6	
8	1	Sample size planning		
8	2	Survey design		
9	1	Midterm presentations 1		
9	2	Midterm presentations 2		

10	1	Preregistration	11
10	2	Meta-analysis	16
11	1	Open science	3
11	2	Visualization 1	15 E
12	1	Visualization 2	
12	2	Exploratory data analysis workshop	
13	1	Sampling, representativeness, and generalizability	4
13	2	Data and participant ethics	12
14	1	Authorship and research ethics	
14	2	[Instructor-specific topics]	
15	1	Open discussion	17
15	2	Open discussion	
16	1	Final project presentations 1	
16	2	Final project presentations 2	

## 8484 A.5.4 16 weeks, content only

Table A.4  
A sample 16-week syllabus with only content materials.

Week	Day	Topic	Chapter
1	1	Class introduction	1
1	2	Theories	2
2	1	Replication and reproducibility	3
2	2	Open science	
3	1	Measurement	8
3	2	Design of experiments 1	9
4	1	Design of experiments 2	
4	2	Sampling	10
5	1	Experimental strategy	
5	2	Preregistration	11
6	1	Data collection	12
6	2	Visualization 1	15
7	1	Visualization 2	
7	2	Midterm exam	
8	1	Introduction to statistics	
8	2	Estimation 1	5
9	1	Estimation 2	
9	2	Inference 1	6

10	1	Inference 2	
10	2	Models 1	7
11	1	Models 2	
11	2	Meta-analysis	16
12	1	Project management	13
12	2	[Instructor-specific topics]	
13	1	[Instructor-specific topics]	
13	2	Sampling, representativeness, and generalizability	4
14	1	Data and participant ethics	
14	2	Authorship and research ethics	
15	1	Ethics: Open discussion	
15	2	Conclusion	17
16	1	Conclusion	
16	2	Final exam	

## 8485 References

Frank, Michael C., and Rebecca Saxe. 2012. "Teaching Replication." *Perspectives on Psychological Science* 7 (6): 600–604. <https://doi.org/10.1177/1745691612460686>.

Hawkins, Robert D., Eric N. Smith, Carolyn Au, Juan Miguel Arias, Rhia Catapano, Eric Hermann, Martin Keil, et al. 2018. "Improving the Replica-

bility of Psychological Science Through Pedagogy.” *Advances in Methods and Practices in Psychological Science* 1 (1): 7–18. <https://doi.org/10.1177/2515245917740427>.

## 8488 B GIT AND GITHUB (ONLINE ONLY)

8489 This appendix appears only in the online version of this book at [https:](https://)

8490 [//experimentology.io/101-github.html](https://experimentology.io/101-github.html).

8491 C R MARKDOWN AND QUARTO (ONLINE

8492 ONLY)

8493 This appendix appears only in the online version of this book [https://  
8494 experimentology.io/102-rmarkdown.html.](https://experimentology.io/102-rmarkdown.html)

8495 **D TIDYVERSE (ONLINE ONLY)**

8496 This appendix appears only in the online version of this book <https://experimentology.io/103-tidyverse.html>.

8497 [experimentology.io/103-tidyverse.html](https://experimentology.io/103-tidyverse.html).

8498 **E GGPLOT (ONLINE ONLY)**

8499 This appendix appears only in the online version of this book <https://experimentology.io/104-ggplot.html>.

8500 [experimentology.io/104-ggplot.html](https://experimentology.io/104-ggplot.html).