

Contents

<i>Preface – How to Read This Book</i>	<i>page xv</i>
<i>Acknowledgments</i>	<i>xvii</i>

PART I INTRODUCTORY TOPICS FOR EVERYONE	1
1 Introduction and Motivation	3
Online Controlled Experiments Terminology	5
Why Experiment? Correlations, Causality, and Trustworthiness	8
Necessary Ingredients for Running Useful Controlled Experiments	10
Tenets	11
Improvements over Time	14
Examples of Interesting Online Controlled Experiments	16
Strategy, Tactics, and Their Relationship to Experiments	20
Additional Reading	24
2 Running and Analyzing Experiments: An End-to-End Example	26
Setting up the Example	26
Hypothesis Testing: Establishing Statistical Significance	29
Designing the Experiment	32
Running the Experiment and Getting Data	34
Interpreting the Results	34
From Results to Decisions	36
3 Twyman's Law and Experimentation Trustworthiness	39
Misinterpretation of the Statistical Results	40
Confidence Intervals	43
Threats to Internal Validity	43
Threats to External Validity	48

	Segment Differences	52
	Simpson's Paradox	55
	Encourage Healthy Skepticism	57
4	Experimentation Platform and Culture	58
	Experimentation Maturity Models	58
	Infrastructure and Tools	66
	PART II SELECTED TOPICS FOR EVERYONE	79
5	Speed Matters: An End-to-End Case Study	81
	Key Assumption: Local Linear Approximation	83
	How to Measure Website Performance	84
	The Slowdown Experiment Design	86
	Impact of Different Page Elements Differs	87
	Extreme Results	89
6	Organizational Metrics	90
	Metrics Taxonomy	90
	Formulating Metrics: Principles and Techniques	94
	Evaluating Metrics	96
	Evolving Metrics	97
	Additional Resources	98
	SIDEBAR: Guardrail Metrics	98
	SIDEBAR: Gameability	100
7	Metrics for Experimentation and the Overall Evaluation Criterion	102
	From Business Metrics to Metrics Appropriate for Experimentation	102
	Combining Key Metrics into an OEC	104
	Example: OEC for E-mail at Amazon	106
	Example: OEC for Bing's Search Engine	108
	Goodhart's Law, Campbell's Law, and the Lucas Critique	109
8	Institutional Memory and Meta-Analysis	111
	What Is Institutional Memory?	111
	Why Is Institutional Memory Useful?	112
9	Ethics in Controlled Experiments	116
	Background	116
	Data Collection	121
	Culture and Processes	122
	SIDEBAR: User Identifiers	123

PART III COMPLEMENTARY AND ALTERNATIVE TECHNIQUES TO CONTROLLED EXPERIMENTS	125
10 Complementary Techniques	127
The Space of Complementary Techniques	127
Logs-based Analysis	128
Human Evaluation	130
User Experience Research (UER)	131
Focus Groups	132
Surveys	132
External Data	133
Putting It All Together	135
11 Observational Causal Studies	137
When Controlled Experiments Are Not Possible	137
Designs for Observational Causal Studies	139
Pitfalls	144
SIDEBAR: Refuted Observational Causal Studies	147
PART IV ADVANCED TOPICS FOR BUILDING AN EXPERIMENTATION PLATFORM	151
12 Client-Side Experiments	153
Differences between Server and Client Side	153
Implications for Experiments	156
Conclusions	161
13 Instrumentation	162
Client-Side vs. Server-Side Instrumentation	162
Processing Logs from Multiple Sources	164
Culture of Instrumentation	165
14 Choosing a Randomization Unit	166
Randomization Unit and Analysis Unit	168
User-level Randomization	169
15 Ramping Experiment Exposure: Trading Off Speed, Quality, and Risk	171
What Is Ramping?	171
SQR Ramping Framework	172
Four Ramp Phases	173
Post Final Ramp	176

16	Scaling Experiment Analyses	177
	Data Processing	177
	Data Computation	178
	Results Summary and Visualization	180
	PART V ADVANCED TOPICS FOR ANALYZING EXPERIMENTS	183
17	The Statistics behind Online Controlled Experiments	185
	Two-Sample t-Test	185
	p-Value and Confidence Interval	186
	Normality Assumption	187
	Type I/II Errors and Power	189
	Bias	191
	Multiple Testing	191
	Fisher's Meta-analysis	192
18	Variance Estimation and Improved Sensitivity: Pitfalls and Solutions	193
	Common Pitfalls	193
	Improving Sensitivity	196
	Variance of Other Statistics	198
19	The A/A Test	200
	Why A/A Tests?	200
	How to Run A/A Tests	205
	When the A/A Test Fails	207
20	Triggering for Improved Sensitivity	209
	Examples of Triggering	209
	A Numerical Example (Kohavi, Longbotham et al. 2009)	212
	Optimal and Conservative Triggering	213
	Overall Treatment Effect	214
	Trustworthy Triggering	215
	Common Pitfalls	216
	Open Questions	217
21	Sample Ratio Mismatch and Other Trust-Related Guardrail Metrics	219
	Sample Ratio Mismatch	219
	Debugging SRMs	222

22	Leakage and Interference between Variants	226
	Examples	227
	Some Practical Solutions	230
	Detecting and Monitoring Interference	234
23	Measuring Long-Term Treatment Effects	235
	What Are Long-Term Effects?	235
	Reasons the Treatment Effect May Differ between Short-Term and Long-Term	236
	Why Measure Long-Term Effects?	238
	Long-Running Experiments	239
	Alternative Methods for Long-Running Experiments	241
	<i>References</i>	246
	<i>Index</i>	266