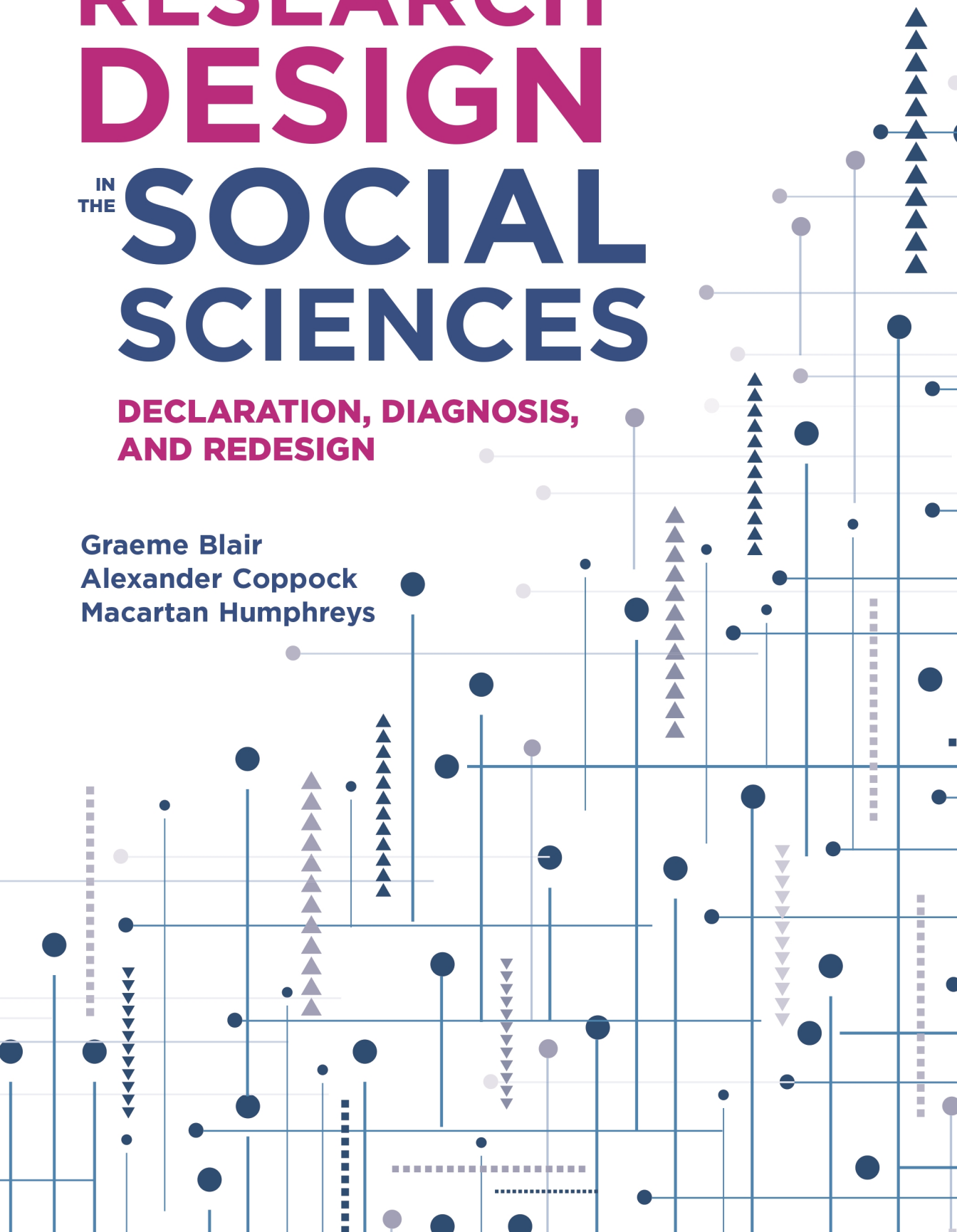
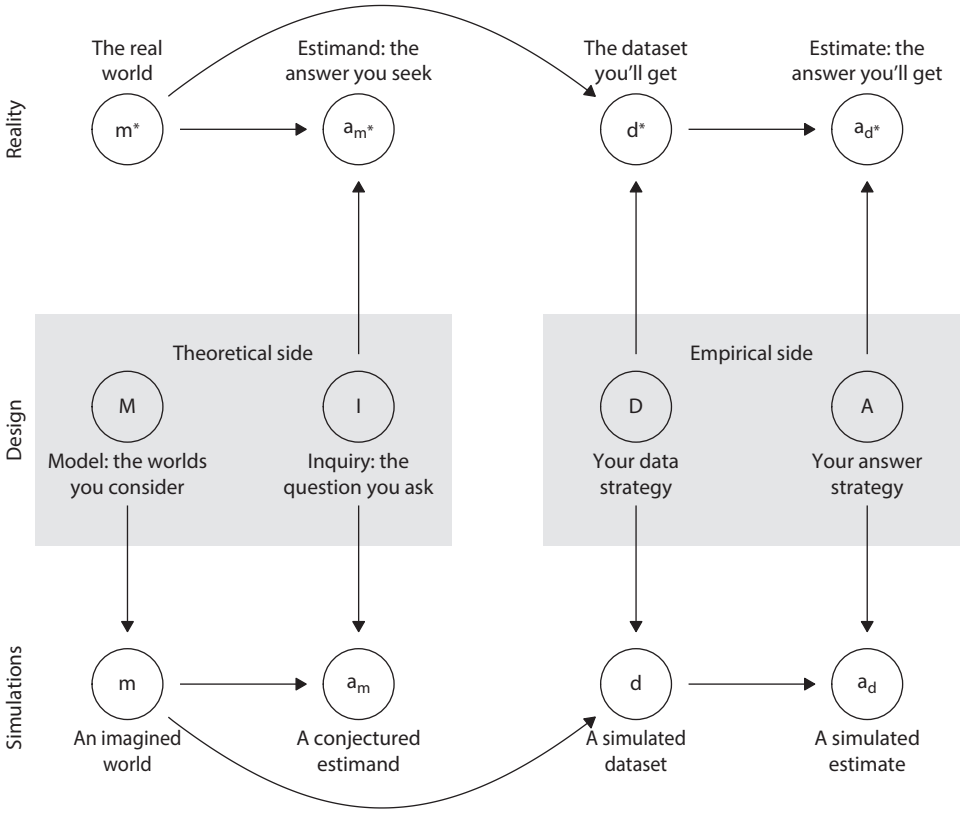
RESEARCH DESIGN IN THE SOCIAL SCIENCES

DECLARATION, DIAGNOSIS, AND REDESIGN

Graeme Blair
Alexander Coppock
Macartan Humphreys



RESEARCH DESIGN in the SOCIAL SCIENCES


Reality
Design
Simulations
The real world
m*
Estimand: the answer you seek
a_{m^*}
The dataset you'll get
d*
Estimate: the answer you'll get
a_{d^*}
Theoretical side
M
Model: the worlds you consider
I
Inquiry: the question you ask
Empirical side
D
Your data strategy
A
Your answer strategy
m
An imagined world
a_m
A conjectured estimand
d
A simulated dataset
a_d
A simulated estimate

RESEARCH DESIGN in the SOCIAL SCIENCES

DECLARATION, DIAGNOSIS, AND REDESIGN

Graeme Blair
Alexander Coppock
Macartan Humphreys

Princeton University Press
Princeton and Oxford

Published by Princeton University Press
41 William Street, Princeton, New Jersey 08540
99 Banbury Road, Oxford OX2 6JX

press.princeton.edu

ISBN 9780691199566
ISBN (pbk.) 9780691199573
ISBN (e-book) 9780691199580

British Library Cataloging-in-Publication Data is available

Editorial: Bridget Flannery-McCoy and Alena Chekanov
Production Editorial: Mark Bellis
Cover Design: Wanda España
Production: Erin Suydam
Publicity: William Pagdatoon
Copyeditor: Bhisham Bherwani

Cover Credit: Burin Suporntawesuk / Alamy Stock Vector

This book has been composed in Minion Pro and Gotham

Printed on acid-free paper. ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgements

We are grateful to have worked with Jasper Cooper for six years on the ideas and tools introduced in this book. His words, ideas, and voice echo throughout the entire project.

Each of us inflicted early versions of `DeclareDesign` and our ideas about research design on our students and colleagues at UCLA, Yale, Columbia, and WZB Berlin, and at many summer schools and workshops. We thank our students for their patience and feedback on the tools and approach, which deeply shaped what you find in this book.

We held a book conference over Zoom and are grateful for the incisive feedback of Dorothy Bishop, Don Green, Nahomi Ichino, Kosuke Imai, Gary King, Andy Gelman, Felix Elwert, Molly Roberts, Cyrus Samii, and Rocio Titiunik. For feedback at our EGAP feedback session on Part I and Part IV, we thank Adam Berinsky, Jake Bowers, David Broockman, Cesi Cruz, Ryan Enos, Alex Hartman, Morgan Holmes, Ryan Moore, Pia Raffler, Dan Rubenson, and Rebecca Wolfe. We thank Abigail Pena-Alejos for her wonderful work constructing the index, Elayne Stecher for work on design examples, Phoenix Dalto for checking code in the book, and Cristian-Liviu Nicolescu and Santiago Sordo Ruz for wonderful last lap support. We also thank the three anonymous reviewers of the manuscript for generous and helpful feedback.

A core part of the `DeclareDesign` project is its software implementation in R. We were lucky to work with a big group of talented graduate students and programmers on nearly every aspect of it. We are grateful to Neal Fultz who moved the software from prototype to professional software product and who made many big contributions to how the tools work now. Luke Sonnet is responsible for the speed and technical workings of estimatr; Aaron Rudkin for many ideas in fabricatr; Clara Bicalho, Markus Konrad, and Sisi Huang for `DeclareDesignWizard`; and Clara Bicalho and Lily Medina for `DesignLibrary`.

We are grateful to the many people who have intensively used the software in its early versions for their generous feedback, Tom Leavitt, Daniel Rubenson, Vartika

Savarna, Tara Slough, Georgiy Syunyaev, Anna Wilke, and Linan Yao. Special thanks also to Dorothy Bishop, Jake Bowers, and Erin Hartman for so many thoughts and suggestions.

We are grateful to the Laura and John Arnold Foundation, and especially Stuart Buck, for seeing the virtues of our approach and providing early funding of the project. We also thank EGAP for seed funding that got us started. We thank Lynn Vavreck for offering Alex a Hoffenberg Visiting Fellowship at UCLA to work on the software and an early version of the manuscript.

We thank Bridget Flannery-McCoy for shepherding this book to fruition and for her willingness to engage with us about offering a free online version of the book. We are also grateful to Eric Crahan for our early conversations about the idea for the book and to Alena Chekanov for guidance in the production phase.

Finally, Graeme and Alex thank Alex's spouse Penelope Van Grinsven for warmly putting up with us for a semester in Los Angeles and a year in New Haven and for many coworking sessions in the pottery studio.

PART

I

Introduction

CHAPTER 1

Preamble

This book introduces a new way of thinking about research designs in the social sciences. Our hope is that this approach will make it easier to develop and to share strong research designs.

At the heart of our approach is the *MIDA* framework, in which a research design is characterized by four elements: a model, an inquiry, a data strategy, and an answer strategy. We have to understand each of the four on their own and also how they interrelate. The design encodes your beliefs about the world, it describes your questions, and it lays out how you go about answering those questions, in terms of both what data you collect and how you analyze it. In strong designs, choices made in the model and inquiry are reflected in the data and answer strategies, and vice versa.

We think of designs as objects that can be interrogated. Each of the four design elements can be "declared" in computer code and—if done right—the information provided is enough to "diagnose" the quality of the design through computer simulation. Researchers can then select the best design for their purposes by "redesigning" over alternative, feasible designs.

This way of thinking pays dividends at multiple points in the research design lifecycle: planning the design, implementing it, and integrating the results into the broader research literature. The declaration, diagnosis, and redesign process informs choices made from the beginning to the end of a research project.

1.1 How to Read This Book

We had multiple audiences in mind when writing this book. First, we were thinking of people looking for a high-level introduction to these ideas. If we only had 30 minutes with a person to try and communicate the core ideas, we would give them Part I. We were thinking of people who are new to the practice of research design and who are embarking on their first empirical projects. The *MIDA* framework introduced in Part I accommodates many different empirical approaches: qualitative and quantitative, descriptive and causal, observational and experimental. Beginners starting out in any of these traditions can use our framework to

consider how the design elements in those approaches fit together. We were also thinking of researchers-in-training: graduate students in seminar courses where the main purpose is to read papers and discuss the credibility of research findings. Such discussions can sometimes feel like a laundry list of complaints, but we hope our framework can focus attention on the most relevant issues. What, exactly, is the inquiry? Is it the right one to be posing? Are the data and answer strategies suited to the inquiry? We were also thinking of funders and decision-makers, who often wish to assess research in terms not of its results but of its design. Our approach provides a way of defining the design and diagnosing its quality.

Part II is more involved. We provide the formal foundations of the *MIDA* framework. We walk through each component of a research design in detail, describe the finer points of design diagnosis, and explain how to carry out a "redesign." We hope Part II will resonate with several audiences of applied researchers both inside and outside of academia. We imagine it could be assigned early in a graduate course on research design in any of the social sciences. We hope data scientists and monitoring and evaluation professionals will find value in our framework for learning about research designs. Scholars will find value in declaring, diagnosing, and redesigning designs whether they are implementing randomized trials or multi-method archival studies, or calibrating structural theories with data.

In Part III, we apply the general framework to specific research designs. The result is a library of common designs. Many empirical research designs are included in the library, but not all. The set of entries covers a large portion of what we see in current empirical practice across social sciences, but it is not meant to be exhaustive.

We are thinking of three kinds of uses for entries in the design library. Collectively, the design entries serve to illustrate the fundamental principles of design. The entries clarify the variety of ways in which models, inquiries, data strategies, and answer strategies can be connected and show how high-level principles operate in common ways across very different designs. The second use is pedagogical. The library entries provide hands-on illustrations of designs in action. A researcher interested in understanding the "regression discontinuity design," for example, can quickly see a complete implementation and learn under what conditions the standard design performs well or poorly. They can also compare the suitability of one type of design against another for a given problem. We emphasize that these descriptions of different designs provide entry points but they are not exhaustive, so we refer readers to recent methodological treatments of the different topics. The third use is as a starter kit to help readers get going on designs of their own. Each entry includes code for a basic design that can be fine-tuned to capture the specificities of particular research settings.

The last section of the book describes how our framework can help at different stages of the research process. Each of these sections should be readable for anyone who

has read Part I. The entry on preanalysis plans, for example, can be assigned in an experiments course as guidance for students filing their first preanalysis plan. The entry on research ethics could be discussed among coauthors at the start of a project. The entry on writing a research paper could be assigned to college seniors writing their first original research papers.

1.2 How to Work This Book

We will often describe research designs not just in words, but in computer code. If you want to work through the code and exercises, fantastic. This path requires investment in R, the `tidyverse`, and the `DeclareDesign` software package. Chapter 4 helps get you started. We think working through the code is very rewarding, but we understand that there is a learning curve. You could tackle the declaration, diagnosis, and redesign processes using any computer language you like,[1] but it is easier in `DeclareDesign` because the software guides you to articulate each of the four design elements.

If you want nothing to do with the code, you can skip it and just focus on the text. We have written the book so that understanding of the code is not required in order to understand research design concepts.

1.3 What This Book Will Not Do

This is a research design book, not a statistics textbook, or a cookbook with recipes applicable to all situations. We will not derive estimators, we will provide no guarantees of the general optimality of designs, and we will present no mathematical proofs. Nor will we provide all the answers to all the practical questions you might have about your design.

What we do offer is a language to express research designs. We can help you learn that language so you can describe your own design in it. When you can declare your design in this language, then you can diagnose it, figure out if it works the way you think it should, and then improve it through redesign.

[1] On our Web site, we provide examples in R, Python, Stata, and Excel.

CHAPTER 2

What Is a Research Design?

At its heart, a research design is a procedure for generating answers to questions. Strong designs yield answers that are close to their targets, but weak designs can produce answers that are misleading, imprecise, or just irrelevant. Assessing whether a design is strong requires having a clear sense of what the question to be answered is and understanding how the empirical information generated or collected by the design will lead to reliable answers. This book offers a language for describing research designs and an algorithm for selecting among them. In other words, it provides a set of tools for characterizing and evaluating the dozens of choices we make in our research activities that together determine the strength of our designs. Throughout, we keep our focus on empirical research designs—designs that seek to answer questions that are answerable with data—and we use the term "research design" as a shorthand for these.

We show that the same basic language can be used to represent research designs whether they target causal or descriptive questions, whether they are focused on theory testing or inductive learning, and whether they use quantitative, qualitative, or a mix of methods. We can select a strong design by applying a simple algorithm: declare-diagnose-redesign. Once a design is declared in simple enough language that a computer can understand it, its properties can be diagnosed through simulation. We can then engage in redesign, or the exploration of a range of neighboring designs. The same language we use to talk to the computer can be used to talk to others. Reviewers, advisers, students, funders, journalists, and the public need to know four basic things to understand a design.

2.1 MIDA: The Four Elements of a Research Design

Research designs share in common that they all have an inquiry *I*, a data strategy *D*, and an answer strategy *A*. Less obviously, perhaps, these three elements presuppose a model *M* of how the world works. We refer to the four together as *MIDA*.

We think of *MIDA* as having two sides. *M* and *I* form the theoretical half, comprising your beliefs about the world and your target of inference. *D* and *A* form the

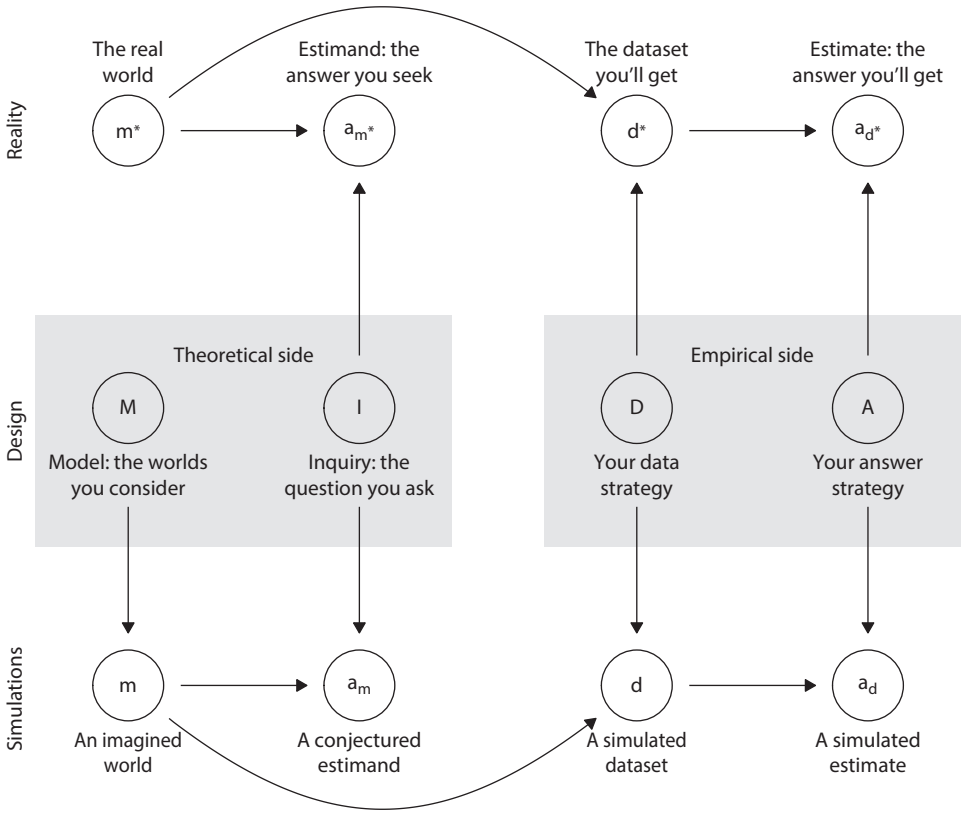


Figure 2.1: The *MIDA* framework. An arrow between two points means that the point at the end of an arrow depends in some way on the one at the start of the arrow. For instance, 'the answer you'll get' depends on the dataset you'll get and the answer strategy you specify.

empirical half, comprising your strategies for collecting and summarizing information. The theoretical side sets the research challenges for you to overcome and the empirical side captures your responses to those challenges.[1]

Figure 2.1 shows how these four elements of a design relate to one another, how they relate to real-world quantities, and how they relate to simulated quantities. We will unpack this figure in the remainder of this chapter, highlighting two especially important parallelisms, first between the upper and lower halves representing actual processes and simulated processes, and second between the left (*M*, *I*) and right (*D*, *A*) halves representing the theoretical and empirical sides of research designs.

[1] We call *M* and *I* the theoretical half because specifying them requires conceptualizing contexts, imagining processes, and posing questions. We call *D* and *A* the empirical half because they describe the empirical strategies. We recognize of course that the theories on the *MI* side may sometimes be very thin and that strategies on the *DA* side should be theoretically motivated.

2.1.1 Model

The set of models in *M* comprises speculations about what causes what and how. It includes guesses about how important variables are generated, how they are correlated, and the sequences of events.

The *M* in *MIDA* does not necessarily represent our beliefs about how the world actually works. Instead, it describes a set of possible worlds in enough detail that we can assess how our design would perform *if* the real world worked like those in *M*. For this reason we sometimes refer to *M* as a set of "reference" models. Assessment of the quality of a design is carried out with reference to the models of the world that we provide in *M*. In other contexts, we might see *M* described as the "data generating process." We prefer to describe *M* as the (imagined) "event generating process" to honor the fact that data are produced or gathered via a data strategy—and the resulting data are measurements taken of the events generated by the world.

We are conscious that the term "model" is used in many different ways by researchers and so a little disambiguation is helpful. Our use of the term when discussing *M*—as a representation of how the world works for the purposes of posing questions and assessing strategies—contrasts with two other usages. First, in some usages, the model is the object of inquiry: our goal in research is to *select* a model of the world that provides a useful representation of the world. We might refer to this as an "inquiry model," to distinguish it from a reference model. We will discuss such approaches and when we do so we will make clear how such models serve a function distinct from *M*. Second, researchers commonly use "model" to describe a representation of event generating processes used specifically for the purpose of generating estimates. For instance, researchers might use a "a linear probability model" or an "ordered probit model." Such "statistical models" might be justified on the grounds that they reflect beliefs about how the world works, but they might also be used simply because they are helpful in generating answers to questions. We think it clearer to think of these models as part of *A*. They are part of the *method* used to answer questions given data. We can then assess, for a given research question, whether the answer strategy provides good answers, whether or not the model assumed by the statistical procedure is consistent with *M*.

2.1.1.1 What's in a model?

The model has two responsibilities. First, the model provides a setting within which a question can be answered. The inquiry *I* should be answerable *under the model*. If the inquiry is the average difference between two possible outcomes, those two outcomes should be described in the model. Second, the model governs what data can be produced by any given data strategy *D*. The data that might be produced by a data strategy *D* should be foreseeable under the model. For example, if the data strategy includes random sampling of units from a population and measurement of

an outcome, the model should describe the outcome variable for all units in that population.

These responsibilities in turn determine what needs to be in the model. In general, the model defines a set of units that we wish to study. Often, this set of units is larger than the set of units that we will actually study empirically, but we can nevertheless define this larger set about which we seek to make inferences. The units might be all of the citizens in Lagos, Nigeria, or every police beat in New Delhi. The set may be restricted to the mayors of cities in California or the catchment areas of schools in rural Poland. The model also includes information about characteristics of those units: how many of each kind of unit there are and how features of the units may be correlated.

For descriptive and causal questions alike, we usually imagine *causal* models. Even if questions are fundamentally descriptive, they can be usefully posed in the context of a causal model, because causal models can explain the level of variables and not simply the nature of effects.

Causal models (see, for instance, Pearl and Mackenzie, 2018) include a set of exogenous and endogenous variables as well as functions that describe the values endogenous variables take depending on the values of other variables. If we think of one variable influencing another, we think of the first as a treatment variable that specifies a condition and the second as an outcome variable. Treatments might be delivered naturally by the world or may be assigned by researchers. The values that an outcome variable would take depending on the level of a treatment are called *potential* outcomes. In the simplest case of a binary treatment, the treated potential outcome is what would arise if the unit were treated, the untreated potential outcome if it were not. Both potential outcomes are part of the model.

Summarizing, we can think of three functions of a model that characterize *units*, *conditions*, and *outcomes*: an identification of a population; a conjecture of values of exogenous variables—conditions; and a description of the values of endogenous variables—outcomes—given the values of other variables on which they depend.

2.1.1.2 M *as a set*

In Figure 2.1, we describe *M* as the "the worlds you'll consider." The reason for this is that we are uncertain about how the world works. As scientists, we are skeptical of easy assertions about what the right model is and we freely admit we don't know the "true model" of the world. Of course the term 'true model' is an oxymoron of sorts, we use it here to highlight the formal similarity between the models in *M* and the true processes we care about. When conducting empirical research into the true model, we have to think through how our design would play out under different possible models, including ones we think more likely and those we think less likely. For instance, the correlation between two variables might be large and

positive, but it could just as well be zero. We might believe that, conditional on some background variables, a treatment has been *as if* randomly assigned by the world—but we might be wrong about that too. In the figure we use m^* to denote the true model, or the actual, unknown, event generating process. We do not have access to m^*, but our hope is that m^* is sufficiently well represented in M so that we can reasonably imagine what will happen when our design is applied in the real world.

How can we construct a sufficiently varied set of models of the world? For this we can draw on existing data from past studies or on new information gathered from pilot studies. Getting a reasonable characterization of the set of plausible models is a core purpose of theoretical reflection, literature review, meta-analysis, and formative research. If there are important known features about your context it generally makes sense to include them in M.

Examples of models

1. Contact theory: When two members of different groups come into contact under specific conditions, they learn more about each other, which reduces prejudice, which in turn reduces discrimination.
2. Prisoner's dilemma. When facing a collective action problem, each of two people will choose noncooperative actions independent of what the other will do.
3. Health intervention with externalities. When individuals receive deworming medication, school attendance rates increase for them and for their neighbors, leading to improved labor market outcomes in the long run.

2.1.2 Inquiry

The inquiry is a research question stated in terms of the model. For example, the inquiry might be the average causal effect of one variable on another, the descriptive distribution of a third variable, or a prediction about the value of a variable in the future. We refer to "the" inquiry when talking about the main research question, but in practice we may seek to learn about many inquiries in a single research study.

Many people use the word "estimand" to refer to an inquiry, and we do too when casually talking about research. When we are formally describing research designs, however, we distinguish between inquiries and estimands, and Figure 2.1 shows why. The inquiry I is the function that operates on the events generated (or conjectured to be generated) by the real world m^* or a simulated world m. The estimand

is the value of that function: a_{m^*} or a_m. In other words, we use "inquiry" to refer to the question and "estimand" to refer to the answer to the question.

As with models, inquiries are also defined with respect to *units*, *conditions*, and *outcomes*: they are summaries of outcomes of units in or across conditions. Inquiries may be causal, as in the sample average treatment effect (SATE). The SATE is the average difference in treated and untreated potential outcomes among units in a sample. Inquiries may also be descriptive, as in a population average of an outcome. While it may seem that descriptive inquiries do not involve conditions, they always do, since the realization of outcomes must take place under a particular set of circumstances, often set by the world and not the researcher.

Figure 2.1 shows that when I is applied to a model m, it produces an answer a^m. This set of relationships forces discipline on both M and I: I needs to be able to return an answer using information available from M and in turn M needs to provide enough information so that I can do its job.

Examples of inquiries

1. What proportion of voters lives with limited exposure to voters from another party in its neighborhood?
2. Does gaining political office make divorce more likely?
3. What types of people will benefit most from a vaccine?

2.1.3 Data strategy

The data strategy is the full set of procedures we use to gather information from the world. The three basic elements of data strategies parallel the three features of inquiries: *units* are selected, *conditions* are assigned, and *outcomes* are measured.

All data strategies require an identification of units. Many involve sampling, gathering data on a subset of units specified by a model or by an inquiry.

Data strategies also involve conditions. Most obviously, experimental interventions are used to produce controlled variation in conditions. If we present some subjects with one piece of information and other subjects with a different piece of information, we've generated variation on the basis of an assignment procedure. Observational approaches often seek to do something similar, selecting units so that natural variation can be exploited. In such cases, units are often selected for study because of the conditions that they are in.

Measurement procedures are the ways in which researchers reduce the complex and multidimensional social world into a parsimonious set of empirical data. These data

need not be quantitative data in the sense of being numbers or values on a predefined scale; qualitative data are data too. Measurement is the vexing but necessary reduction of reality to a few choice representations.

Figure 2.1 shows how the data strategy is applied to both the imagined worlds in M and to the real world. When D is applied to the real world (m^*), we obtain the realized dataset d^*. When D is applied to the worlds we imagine in M, we obtain *simulated* datasets, which may or may not be like the dataset d^* we would really get. When our models M more accurately represent the real world, our simulated datasets will look more like the real data we will eventually collect.

Examples of data strategies

Sampling procedures.

1. Random digit dial sampling of 500 voters in the Netherlands
2. Respondent-driven sampling of people who are HIV positive, starting from a sample of HIV-positive individuals
3. "Mall intercept" convenience sampling of men and women present at the mall on a Saturday

Treatment assignment procedures.

4. Random assignment of free legal assistance intervention for detainees held in pretrial detention
5. Nature's assignment of the sex of a child at birth

Measurement procedures.

6. Voting behavior gathered from survey responses
7. Administrative data indicating voter registration
8. Measurement of stress using cortisol readings

2.1.4 Answer strategy

The answer strategy is what we use to summarize the data produced by the data strategy. Just like the inquiry summarizes a part of the model, the answer strategy summarizes a part of the data. We can't just "let the data speak" because complex, multidimensional datasets don't speak for themselves—they need to be summarized and explained. Answer strategies are the procedures we follow to do so.

Answer strategies are functions that take in data and return answers. For some research designs, this is a literal function like the R function `lm_robust` that implements an ordinary least squares (OLS) regression with robust standard errors. For some research designs, the function is embodied by the researchers themselves when they read documents and summarize their meanings in a case study.

The answer strategy is more than the choice of an estimator. It includes the full set of procedures that begins with cleaning the dataset and ends with answers in words, tables, and graphs. These activities include data cleaning, data transformation, estimation, plotting, and interpretation. Not only do we define our choice of OLS as the estimator, we also specify that we will focus attention on a particular coefficient estimate, assess uncertainty using a 95% confidence interval, and construct a coefficient plot to visualize the inference. The answer strategy also includes all of the if-then procedures that researchers implicitly or explicitly follow depending on initial results and features of the data. For example, in a stepwise regression procedure, the answer strategy is not the final regression specification that results from iterative model selection, but the whole procedure.

D and *A* impose a discipline on each other in the same way as we saw with *M* and *I*. Just as the model needs to provide the events that are summarized by the inquiry, the data strategy needs to provide the data that are summarized by the answer strategy. Declaring each of these parts in detail reveals the dependencies across the design elements.

A and *I* also enjoy a tight connection stemming from the more general parallelism between (*M*, *I*) and (*D*, *A*). We elaborate the principle of parallel inquiries and answer strategies in Section 9.3.

Figure 2.1 shows how the same answer strategy *A* is applied both to the realized data d^* and to the simulated data *d*. We know that in practice, however, the *A* applied to the real data differs somewhat from the *A* applied to the data we plan for via simulation. Designs sometimes drift in response to data, but too much drift and the inferences we draw can become misleading. The *MIDA* framework encourages researchers to think through what the real data will actually look like, and adjust *A* accordingly *before* data strategies are implemented.

Examples of answer strategies

1. Multilevel modeling and poststratification
2. Bayesian process tracing
3. Difference-in-means estimation

2.2 Declaration, Diagnosis, Redesign

With the core elements of a design described, we are now ready to lay out the declaration, diagnosis, and redesign workflow.

2.2.1 Declaration

Declaring a design entails figuring out which parts of your design belong in *M*, *I*, *D*, and *A*. The declaration process can be a challenge because mapping our ideas about a project into *MIDA* is not always straightforward, but it is rewarding. When we can express a research design in terms of these four components, we are newly able to think about its properties.

Designs can be declared in words, but declarations often become much more specific when carried out in code. You can declare a design in any statistical programming language: Stata, R, Python, Julia, SPSS, SAS, Mathematica, among many others. Design declaration is even possible—though somewhat awkward—in Excel. We wrote the companion software, `DeclareDesign`, in R because of the availability of other useful tools in R and because it is free, open-source, and high-quality. We have designed the book so that you can read it even if you do not use R, but you will have to translate the code into your own language of choice. On our Web site, we have pointers for how you might declare designs in Stata, Python, and Excel. In addition, we link to a "Design wizard" that lets you declare and diagnose variations of standard designs via a point-and-click Web interface. Chapter 4 provides an introduction to `DeclareDesign` in R.

2.2.2 Diagnosis

Once you've declared your design, you can diagnose it. Design diagnosis is the process of simulating a research design in order to understand the range of ways the study could turn out. Each run of the design comes out differently because different units are sampled, or the randomization allocates different units to treatment, or outcomes are measured with different errors. We let computers do the simulations for us because imagining the full set of possibilities is—to put it mildly—cognitively demanding.

Diagnosis is the process of assessing the properties of designs, and provides an opportunity to write down what would make the study a success. For a long time, researchers have classified studies as successful or not based on statistical significance (Chopra et al., 2022). If significant, the study "worked"; if not, it is a failed "null." Accordingly, statistical power (the probability of a statistically significant result) has been the most front-of-mind design property when researchers plan studies. As we learn more about the pathologies of relying on statistical significance, we learn that features beyond power are more important. For example, the

"credibility revolution" throughout the social sciences has trained a laser-like focus on the biases that may result from omitted or "lurking" variables.

Design diagnosis relies on two new concepts: diagnostic statistics and diagnosands.

A "diagnostic statistic" is a summary statistic generated from a single "run" of a design. For example, the statistic *e* (error) refers to the difference between the estimate and the estimand. The statistic *s* (significance) refers to whether the estimate was deemed statistically significant at the 0.05 level (for instance).

A "diagnosand" is a summary of the distribution of a diagnostic statistic across many simulations of the design. The bias diagnosand is defined as the average value of the *e* statistic and the power diagnosand is defined as the average value of the *s* statistic. Other diagnosands include quantities like root-mean-squared error (RMSE), Type I and Type II error rates, how likely it is that subjects were harmed, and average cost. We describe these diagnosands in much more detail in Chapter 12.3.

One especially important diagnosand is the "success rate," which is the average value of the "success" diagnostic statistic. As the researcher, you get to decide what would make your study a success. What matters most in your research scenario? Is it statistical significance? If so, optimize your design with respect to power. Is what matters most whether the answer has the correct sign or not? Then diagnose how frequently your answer strategy yields an answer with the same sign as your estimand. Diagnosis involves articulating what would make your study a success and then figuring out, through simulation, how likely you are to obtain that success. Success is often a multidimensional aggregation of diagnosands, such as the joint achievement of high statistical power, manageable costs, and low ethical harms.

We diagnose studies over the range of possibilities in the model, since we want to learn the value of diagnosands under many possible scenarios. A clear example of this is the power diagnosand over many possible conjectures about the true effect size. For each effect size that we entertain in the model, we can calculate statistical power. The minimum detectable effect size is a summary of this power curve, usually defined as the smallest effect size at which the design reaches 80% statistical power. This idea, however, extends well beyond power. Whatever the set of important diagnosands, we want to ensure that our design performs well across many model possibilities.

Computer simulation is not the only way to do design diagnosis. Designs can be declared in writing or mathematical notation and then diagnosed using analytic formulas. Enormous theoretical progress in the study of research design has been made with this approach. Methodologists across the social sciences have described diagnosands such as bias, power, and root-mean-squared error for large classes of designs. Not only can this work provide closed-form mathematical expressions for many diagnosands, it can also yield insights about the pitfalls to watch out for when

constructing similar designs. That said, pen-and-paper diagnosis is challenging for many social science research designs, first because many designs—as actually implemented—have idiosyncratic features that are hard to incorporate and, second, because the analytic formulas for many diagnosands have not yet been worked out by statisticians. For these reasons, when we do diagnosis in this book we will usually depend on simulation.

Even when using simulation, design diagnosis doesn't solve every problem and, like any tool, it can be misused. We outline two main concerns. The first is the worry that the diagnoses are plain wrong. Given that design declaration includes conjectures about the world, it is possible to choose inputs such that a design passes any diagnostic test set for it. For instance, a simulation-based claim of unbiasedness that incorporates all features of a design is still only good with respect to the precise conditions of the simulation. In contrast, analytic results, when available, may extend over general classes of designs. Still worse, simulation parameters might be chosen opportunistically. Power analysis is useless if implausible parameters are chosen to raise power artificially. While our framework may encourage more principled declarations, it does not guarantee good practice. As ever, garbage-in, garbage-out. The second concern is the risk that research may be evaluated on the basis of a narrow or inappropriate set of diagnosands. Statistical power is often invoked as a key design feature, but well-powered studies that are biased are of little use. The importance of particular diagnosands can depend on the values of others in complex ways, so researchers should take care to evaluate their studies along many dimensions.

2.2.3 Redesign

Once your design has been declared, and you have diagnosed it with respect to the most important diagnosands, the last step is redesign.

Redesign entails fine-tuning features of the data and answer strategies to understand how they change your diagnosands. Most diagnosands depend on features of the data strategy. We can redesign the study by varying the sample size to determine how big it needs to be to achieve a target diagnosand: 90% power, say, or an RMSE of 0.02. We could also vary an aspect of the answer strategy, such as the choice of covariates used to adjust a regression model. Sometimes the changes to the data and answer strategies interact. For example, if we want to use covariates that increase the precision of the estimates in the answer strategy, we have to collect that information as a part of the data strategy. The redesign question now becomes, is it better to collect pretreatment information from all subjects or is the money better spent on increasing the total number of subjects and only measuring post-treatment?

The redesign process is mainly about optimizing research designs given ethical, logistical, and financial constraints. If diagnosands such as total harm to subjects,

total researcher hours, or total project cost exceed acceptable levels, the design is not feasible. We want to choose the best design we can among the feasible set. If the designs remaining in the feasible set are underpowered, biased, or are otherwise scientifically inadequate, the project may need to be abandoned.

In our experience, it's during the redesign process that designs become *simpler*. We learn that our experiment has too many arms or that the expected level of heterogeneity is too small to be detected by our design. We learn that in our theoretical excitement, we've built a design with too many bells and too many whistles. Some of the complexity needs to be cut, or the whole design will be a muddle. The upshot of many redesign sessions is that our designs pose fewer questions but obtain better answers.

2.3 Example: A Decision Problem

Imagine you want to study whether a new policy—implicit bias training—changes social norms of police officers or is merely window dressing. You have a research budget of $3,000 to run a randomized experiment to test the training program. You expect the police department will scale up the training program across the force if you find it shifts norms by at least 0.3 standard units, and otherwise it will not be implemented more widely. The department is also enamored by classical statistical testing so they will likely only go forward if your estimates are statistically significant.

Though we describe this particular setting to fix ideas, we think this example is relevant for many decision problems in which the results of a study will inform implementation.

You will consider the experiment to be a success if you conclude that the program is effective and indeed it is effective (which in this example we will take to mean that there is in fact an effect of at least 0.2). Otherwise you consider it a failure, whether because you reached the wrong conclusion or because resources were spent that could have been used on an effective intervention.

For the experiment itself, you're deciding between two designs. In one you run a study with 150 officers, randomly assign half to receive the training and half to not receive it, then compare outcomes in treatment and control using a survey about their perceived norms of reporting. In the second, you spend part of your funding gathering background information on the officers—whether they have been investigated in the past by internal affairs and were found to have discriminated against citizens—and use that information to improve both randomization and inference. Let's suppose the two designs cost exactly the same amount. Interviewing each officer at endline costs $20, so the total cost of the larger trial is $150 * 20 = 3{,}000$. The block-randomized design costs the same for endline measurement, but

measurement of the history variable from police administrative records costs $10 per individual because you have to go through the department's archives, which are not digitized, so the total is the same: $100 * 20 + 100 * 10 = 3{,}000$.

The two designs cost the same but differ on the empirical side. Which strategy should you use, given your goals?

2.3.1 Design 1: N = 150, complete random assignment

- *M*: We first define a model that stipulates a set of 18,000 units representing each officer and an unknown treatment effect of the training lying somewhere between 0 and 0.5. This range of possible effects implies that in 60% of the models we consider, the true effect is above our threshold for a program worth implementing, 0.2. Outcomes for each individual depend on their past infractions against citizens (their history). The importance of history is captured by the parameter b. We don't know how important the history variable is, so we will simulate over a plausible range for b. *M* here is a *set* of models as each "run" of the model will presuppose a different treatment effect for all subjects as well as distinct outcomes for all individuals.
- *I*: The inquiry is the difference between the average treated outcome and the average untreated outcome, which corresponds to the average treatment effect. We are writing it this way to highlight the similarity between the inquiry and the difference-in-means answer strategy that we will adopt.
- *D*: We imagine a data strategy with three components relating to units, conditions, and outcomes: we sample 100 deputies to participate in the experiment, assign exactly half to treatment and the remainder to control, and finally measure their outcomes through a survey.
- *A*: The answer strategy takes the difference-in-means between the treated and untreated units. Thus the answer strategy uses a function similar to the inquiry itself.

When we put these all together we have a design, Declaration 2.1.

Declaration 2.1 Two-arm trial design.

```
b <- 0

model <-
  declare_model(
    N = 1000,
    history = sample(c(0, 1), N, replace = TRUE),
    potential_outcomes(Y ~ b * history + runif(1, 0, 0.5) * Z + rnorm(N)))
```

```
inquiry <-
  declare_inquiry(ATE = mean(Y_Z_1) - mean(Y_Z_0))

data_strategy <-
  declare_sampling(S = complete_rs(N = N, n = 150), filter = S == 1) +
  declare_assignment(Z = complete_ra(N)) +
  declare_measurement(Y = reveal_outcomes(Y ~ Z))

answer_strategy <-
  declare_estimator(Y ~ Z, .method = difference_in_means, inquiry = "ATE")

declaration_2.1 <- model + inquiry + data_strategy + answer_strategy
```

Table 2.1: Simulated data from two-arm trial design.

ID	history	Y_Z_0	Y_Z_1	S	Z	Y
0003	0	−2.01	1.30	1	1	1.30
0015	0	1.33	0.77	1	1	0.77
0017	1	−1.07	−0.86	1	1	−0.86
0021	0	−0.39	3.09	1	0	−0.39
0024	0	1.01	1.45	1	1	1.45
0034	1	−0.69	0.77	1	1	0.77

The design is now ready to be used, diagnosed, developed. We can generate simulated data directly from the design using `draw_data(declaration_2.1)`. We show a snapshot of such simulated data below.

To evaluate the design, we need to specify our criteria for what counts as a good design. We could assess the design in terms of its statistical power, whether estimation is unbiased, and so on. For now though we will focus on a specific design characteristic, its "success rate," which is the probability you will deem the research a success, using the criteria defined above.[2]

We specify the criteria for success in this call to `declare_diagnosands`:

```
program_diagnosands  <-
  declare_diagnosands(
    success = mean(estimate > 0.3 & p.value < 0.05 & estimand > 0.2)
  )
```

[2]We could define more complex diagnosands that, for example, give correct decisions to implement a positive weight and incorrect decisions to implement a negative weight. The diagnosands you choose should reflect what you care about most in any given design setting.

2.3.2 Design 2: N = 100, baseline measurement, block random assignment

The alternative design differs on the empirical side in three ways. First, fewer subjects are sampled. Second, information about the subjects' background (their "history") is used to implement a block randomization that conditions assignment on history. Third, the subjects' history is taken into account in the analysis. This last choice is an instance of adjusting the answer strategy in light of a change to the data strategy.

In Declaration 2.2, we can leave the model and inquiry intact, but we have to work on the data and answer strategies.

Declaration 2.2 A design that exploits background information.

```
data_strategy_2 <-
  declare_sampling(S = complete_rs(N = N, n = 100),
                   filter = S == 1) +
  declare_assignment(Z = block_ra(blocks = history)) +
  declare_measurement(Y = reveal_outcomes(Y ~ Z))

answer_strategy_2 <-
  declare_estimator(Y ~ Z, .method = difference_in_means,
                    blocks = history, inquiry = "ATE")

declaration_2.2 <-
  model + inquiry + data_strategy_2 + answer_strategy_2
```

2.3.3 Diagnosis and comparison

We can then diagnose both designs over a series of conjectured values for the importance of history (`b`) and see how they perform on our specified criterion for success.

Diagnosis 2.1 Diagnosis of `declaration_2.1` and `declaration_2.2`.

```
declaration_2.1 |>
  redesign(b = seq(0,3,0.25)) |>
  diagnose_design(diagnosands = program_diagnosands)

declaration_2.2 |>
  redesign(b = seq(0,3,0.25)) |>
  diagnose_design(diagnosands = program_diagnosands)
```

The results are shown in Figure 2.2.

When background factors don't make much of a difference for the social norms outcome, the first design outperforms the second: after all, the first design has a sample size of 150 compared with the second design's 100. We're successful over

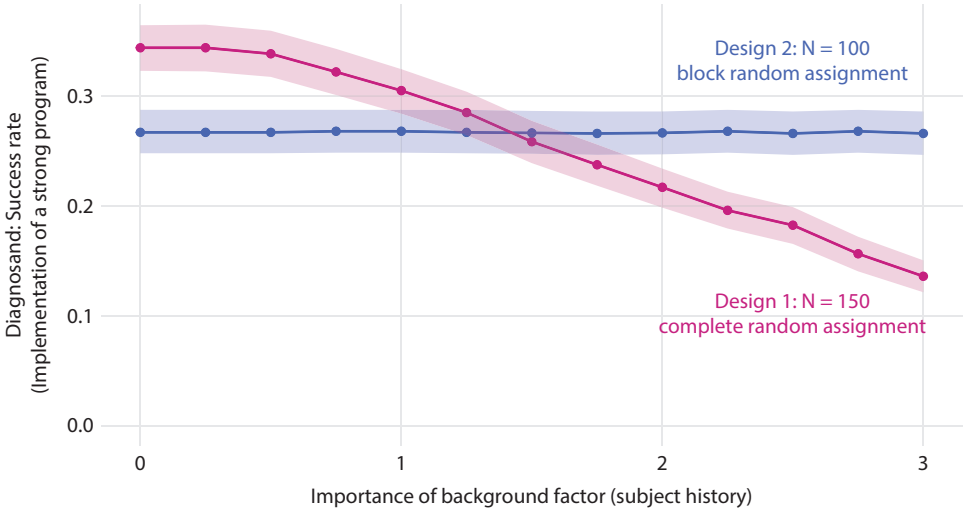


Figure 2.2: How success depends on choice of D and A given different possibilities for M.

30% of the time when using the first design, compared with about 25% when using the second. These rates seem low, but recall that the treatment effect variation we built into the model implies that the program is *worth* implementing only 60% of the time, because the other 40% of the time the true effects are smaller than 0.2.

As subject history has a bigger impact on the outcome variable, however, the first design does worse and worse. In essence, the additional variation due to background factors makes it more difficult to separate signal from noise, making it more likely that our estimates are not significant and therefore more likely that we decline to implement the program.

Here is where the smaller design that blocks on subject history shines: this variation is conditioned in two places, in the assignment strategy and in the estimator. The result is a more precise procedure that is better able to separate signal from noise. Ultimately, the blocked design has the same success rate regardless of the importance of the background factors.

The overall result of this declaration, diagnosis, and redesign process is that which design you choose depends on *beliefs* about the importance of background conditions for outcomes. Now the design question hinges on something you can go learn about: how much variation is explained by subject history?

2.3.4 Three principles

We see from this example the gains from entertaining a diverse model set rather than presupposing we already know M. We also see an example of design parts tailored to each other, most importantly the adjustment of answer strategies in light

of data strategies. And we see that design choices are informed by a clear specification of a success criterion. In the next chapter we develop these three features as broader principles, referring to them as Principle 3.1: *Design holistically*, Principle 3.2: *Design agnostically*, and Principle 3.3: *Design for purpose*.

2.4 Putting Designs to Use

The two pillars of our approach are the language for describing research designs (*MIDA*) and the algorithm for selecting high-quality designs (declare, diagnose, redesign). Together, these two ideas can shape research design decisions throughout the lifecycle of a project. The full set of implications is drawn out in Part IV but we emphasize the most important ones here.

Broadly speaking, the lifecycle of an empirical research project has three phases: planning, realization, and integration. Having a clear characterization of your design in terms of *MIDA* is helpful in all three of these stages.

2.4.1 Planning, realization, integration

Planning entails some or all of the following steps, depending on the design: conducting an ethical review, seeking human subjects approval, gathering criticism from colleagues and mentors, running pilot studies, and preparing preanalysis documents. The design as encapsulated by *MIDA* will go through many iterations and refinements during this period, but the goal is simple: to assess whether your data strategy and answer strategy are capable of providing reliable answers to your inquiry given different models that you might entertain. Planning is the time when frequent reapplication of the declare, diagnose, redesign algorithm will pay the highest dividends. How should we investigate the ethics of a study? Consider casting the ethical costs and benefits as diagnosands. How should we respond to criticism, constructive or not? By reinterpreting the feedback in terms of *M*, *I*, *D*, and *A*. How can we convince funders and partners that our research project is worth investing in? By credibly communicating our study's diagnosands: its statistical power, its unbiasedness, and its high chance of success, however the partner or funder defines it. What belongs in a preanalysis plan? You guessed it—a specification of the model, inquiry, data strategy, and answer strategy.

Realization is the phase of research in which all those plans are executed. We implement the data strategy in order to gather information from the world. Once that's done, we follow the answer strategy in order to finally generate answers to the inquiry. Of course, that's only if things go exactly according to plan, which they never do. Survey questions don't work as we imagine, partner organizations lose interest in our study, subjects move or become otherwise unreachable. A critic or a reviewer may insist we change our answer strategy, or may think a different inquiry

altogether is theoretically more appropriate. We may ourselves change how we think of the design as we embark on writing up the research project. It is likely that some features of *MIDA* will change during the realization phase, in which case you can again use diagnosis to assess whether changes to *MIDA* are for good or for bad. Some design changes have very bad properties, like sifting through the data ex-post, finding a statistically significant result, then backfitting a new I to match the new A. Indeed, if we declare and diagnose this actual answer strategy (sifting through data ex-post), we can show through design diagnosis that it yields misleading answers. Other changes made along the way may help the design quite a bit. If the planned design did not include covariate adjustment, but a friendly critic suggests adjusting for the pretreatment measure of the outcome, the "standard error" diagnosand might drop nicely. The point here is that design changes during the implementation process, whether necessitated by unforeseen logistical constraints or required by the review process, can be understood in terms of M, I, D, and A by reconciling the planned design with the design as implemented.

A happy realization phase concludes with the publication of results. But the research design lifecycle is not finished: the study and its results should be integrated into the broader community of scientists, decision-makers, and the public. Studies should be archived, along with design information, to prepare for reanalysis. Future scholars may well want to reanalyze your data in order to learn more than is represented in the published article or book. Good reanalysis of study data requires a full understanding of the design as implemented, so archiving design information along with code and data is critical. Not only may your design be reanalyzed, it may also be replicated with fresh data. Ensuring that replication studies answer the same theoretical questions as original studies requires explicit design information, without which replicators and original study authors may simply talk past one another. Indeed, as our studies are integrated into the scientific literature and beyond, we should anticipate disagreement over our claims. Resolving disputes is very difficult if parties do not share a common understanding of the research design. We might also anticipate that our results will be formally synthesized with others' work via meta-analysis. Meta-analysts need design information in order to be sure they aren't inappropriately mixing together studies that ask different questions or answer them too poorly to be of use. Finally, with luck your designs will be a model for others. Having an analytically complete representation of your design at hand will make it that much easier to use redesign to build on what you have done.

2.4.2 Three more principles

This discussion motivates three more principles: Principle 3.4: *Design early* to reap the benefits of clarity; Principle 3.5: *Design often* so that you can correct course; and Principle 3.6: *Design to share* so that you maximize transparency and contribute maximally to knowledge creation.

CHAPTER 3

Research Design Principles

With the *MIDA* framework and the declare, diagnose, redesign algorithm in hand, we can articulate a set of six principles for research design.

This section offers succinct discussions of each principle. We will expand on the implications of these principles for specific design choices throughout the book.

Design principles

1. Design holistically
2. Design agnostically
3. Design for purpose
4. Design early
5. Design often
6. Design to share

Principle 3.1 Design holistically

This is perhaps the most important of our principles. Designs are good not because they have good components but because the components work together to get a good result. Too often, researchers develop and evaluate parts of their designs in isolation: Is this a good question? Is this a good estimator? What's the best way to sample? But if you design with a view to diagnosis you are forced to focus on how each part of the design fits together. An estimator might be appropriate if you use one assignment scheme but not another. The evaluation of data and answer strategies depends on whether your model and inquiry call for descriptive inference, causal inference, or generalization inference (or perhaps, all three at once). If we ask, "What's your research design?" and you respond "It's a regression discontinuity design," we've learned something about what class your answer strategy might fall into, but we don't have enough information to decide whether it's a strong design

until we learn about the model, inquiry, data strategy, and other parts of the answer strategy. Ultimately design evaluation comes not from assessment of the parts but from diagnosis of the full design.

When we consider whole designs rather than just thinking about one aspect at a time, we notice how designs that have "parallel" theoretical and empirical sides tend to be strong. We develop this idea in Section 9.3. If you want your estimate $a_{d^*} = A(d)$ to be close to the estimand $a_{m^*} = I(m^*)$, it's often best to choose data strategies that parallel models and answer strategies that parallel inquiries, i.e., to make sure that this rough analogy holds: $M{:}I{::}D{:}A$.

Principle 3.2 Design agnostically

When we design a research study, we have in mind a model of how the world works. But a good design should work, and work well, even when the world is different from what we expect. One implication is that we should entertain many models, not just seeking to ensure the design produces good results for models that we think likely but also trying to expand the set of possible models for which the design delivers good results. A second implication is that inquiries and answer strategies should still *work* when the world looks different from what we expect. Inquiries should have answers even when event generating processes are different from how you imagine them. In the same way, the ability to apply an answer strategy should depend as little as possible on strong expectations of how the data you will get will look.

A corollary to "Design agnostically" is that we should know for which models our design performs well and for which models it performs poorly. We want to diagnose over many models to find where designs break. All designs break under some models, so the fact that a design ever breaks is no criticism. As research designers, we just want to know which models pose problems and which do not.

Principle 3.3 Design for purpose

When we say a design is good we mean it is good for some specific purpose. That purpose should be captured by the diagnosands used to assess design quality and design decisions should then be taken with respect to the specified purpose. Too often, researchers focus on a narrow set of diagnosands, and consider them in isolation. Is the estimator unbiased? Do I have statistical power? The evaluation of a design nearly always requires balancing multiple criteria: scientific precision, logistical constraints, policy goals, as well as ethical considerations. And oftentimes these might come into conflict with each other. Thus one design might be best if the goal is to assess whether a treatment has any effect, another if the goal is to assess the size of an effect. One design might be optimal if the goal is to contribute to general knowledge about how processes work, but another if the goal is to make a decision about whether to move forward with a policy in a given context.

In the MIDA framework, the goals of a design are not formally a part of the design. They enter at the diagnosis stage, and, of course, a single design might be assessed for performance for different purposes.

Principle 3.4 Design early

Designing an empirical project entails declaring, diagnosing, and redesigning the components of a research design: its model, inquiry, data strategy, and answer strategy. The design phase yields the biggest gains when we design early. By front-loading design decisions, we can learn about the properties of a design while there is still time to improve them. Once data strategies are implemented—units sampled, treatments assigned, and outcomes measured—there's no going back. While applying the answer strategy to the revealed dataset, you might well wish you'd gathered data differently, or asked different questions. Post-hoc, we always wish our previous selves had planned ahead.

A reason deeper than regret for designing early is that the declaration, diagnosis, and redesign process inevitably changes designs, almost always for the better. Revealing how each of the four design elements are interconnected yields improvements to each. These choices are almost always better made before any data are collected or analyzed.

Principle 3.5 Design often

Designing early does not mean being inflexible. In practice, unforeseen circumstances may change the set of feasible data and answer strategies. Implementation failures due to nonresponse, noncompliance, spillovers, inability to link datasets, funding contractions, or logistical errors are common ways the set of feasible designs might contract. The set of feasible designs might expand if new data sources are discovered, additional funding is secured, or if you learn about a new piece of software. Whether the set expands or contracts, we benefit from declaring, diagnosing, and redesigning given the new realities.

In part IV on the research design lifecycle, we push this principle to the limit, encouraging you to keep on designing even after research is completed, arguing that *ex post* design can help you assess the robustness of your claims and help you decide how to respond to criticism of your work.

Principle 3.6 Design to share

The *MIDA* framework and the declaration, diagnosis, and redesign algorithm can improve the quality of your research designs. It can also help you communicate your work, justify your decisions, and contribute to the scientific enterprise. Formalizing design declaration makes this sharing easier. By coding up a design as an object that can be run, diagnosed, and redesigned, you help other researchers see, understand, and question the logic of your research.

We urge you to keep this sharing function in mind as you write code, explore alternatives, and optimize over designs. An answer strategy that is hard-coded to capture your final decisions might break when researchers try to modify parts. Alternatively, designs can be created specifically to make it easier to explore neighboring designs, let others see why you chose the design you chose, and give them a leg up in their own work. In our ideal world, when you create a design, you contribute it to a design library so others can check it out and build on your good work.