Rolf Steyer

Probability and Causality

Version: April 11, 2016

University of Jena

Preface

What can we do to reduce global warming? How can we prevent another global financial crisis? How to fight AIDS? How can we reduce hunger in the world? These questions ask about causal effects of interventions. Obviously, interventions based on the wrong causal theories and hypotheses will cost the life of many and huge amounts of money that could be spent more appropriately. Even if our daily problems are less dramatic, they are of the same nature. Just think about your own actions that you have to chose in your responsibilities as a student, scientist, teacher, physician, psychologist, politician, or just as a parent! Whatever you do has effects, and these effects might be different if you take one action instead of another one. It is these kind of thoughts that make us believe that there is no other issue in the methodology of empirical sciences that deserves and needs more attention and effort than causality. And because the dependencies we are investigating are of a nondeterministic nature, we need a *probabilistic theory of causality*. In other words, we need to understand *probability* and *causality*.

What This Book is About

Empirical causal research involves several inferences and interpretations. Among these are:

(a) statistical inference, i. e., the inference from sample data to parameters characterizing the distributions of random variables,
(b) causal inference, i. e., the inference from parameters characterizing the distributions of random variables to causal effects and/or dependencies,
(c) interpretation of the putative cause,
(d) interpretation of the outcome variable,
(e) interpretation of the random experiment considered.

This book does not deal with all these points. We will neither discuss the mathematics of statistical inference nor the content issues of construct validity or external validity (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) involved in points (c) to (e). Instead we will focus on the second point: causal inference, i. e., the inference from parameters (such as the expectations of an outcome variable in two treatment conditions) to causal effects and/or causal dependencies. This is what the probabilistic theory of causality presented in this book is about. As will be shown in this book, causal effects are

also parameters that characterize the joint distributions of the random variables considered in a random experiment. However, their definitions are less obvious than 'ordinary' expectations and their differences.

Basic Idea

In order to get a first impression of what this means, let us briefly formulate the basic idea that can most easily be explained if the putative (or presumed) cause is a treatment variable. Suppose an individual, or in more general terms, an observational unit, could be treated by condition 1 or it could be treated by condition 0, *everything else being invariant*. If there is a difference in the outcome considered (some measure of success of the treatment), then this difference is due to the difference in the two treatment conditions. This conception goes back at least to Mill (1843/1865).

Multiple Determinacy

The problem with this first version of the basic idea is that most outcomes are *multiply determined*, i. e., they are not only influenced by the treatment variable, but by many other variables as well. In the field of agricultural research, e. g., the *yield* (*outcome*) of a *variety* not only depends on the variety (*treatment*) itself, but it also depends on the quality of the *plot* (*observational unit*), such as the average hours of sunshine on the plot per day, the amount of water reaching the plot, and the number of microbes in the plot, etc. Although Mill's idea sounds perfect, it is not immediately clear which implications it has for practice, because the number of other causes is often too large for keeping constant all of them. Furthermore Mill's idea fails to distinguish between covariates and intermediate variables. Holding constant all intermediate as well — and not only all covariates — would imply that there is no treatment effect any more, if we assume that all treatment effects have to be transmitted by some intermediate(s) (see section 3.3 for a more detailed discussion).

Because of the problem of multiple determinacy, Mills conception has been complemented by Sir Ronald A. Fisher (1925/1946) and by Jerzy S. Neyman (1923/1990) in the second and third decades of the last century. Simply speaking, introducing the randomized experiment, Fisher replaced the *ceteris paribus* clause ('everything else invariant') by the *ceteris paribus distributionibus* clause: *all other possible causes (the 'covariates') having the same distribution*. This is what random assignment of units to treatment conditions secures.

A Metaphor — The Invisible Man and his Shadow

Imagine an invisible man. Although we cannot see him, suppose we know that he is there, because we can see his shadow. Furthermore, suppose we would like to measure his size. Doing that, we have two problems, a theoretical and a practical one. The *theoretical problem* is to define *size*. We have to clarify that we do not

mean 'volume' or 'weight', but 'height' — without shoes, and without hat and hair. Unfortunately, actual height varies slightly in the course of a day. Hence, we define *size* to be the average of the actual heights at the different times of the day. This solves the theoretical problem; now we know what we want to measure.

However, because the man is invisible, we cannot measure his *size* directly — and this is not only because his size slightly varies over the day. The crucial problem is that we can only observe his shadow. And this is the *practical problem*: How to determine his size from his shadow? Sometimes, there is almost no shadow at all, sometimes it is huge. Some geometrical reflection yields a first simple solution: measuring the shadow when the sun has an angle of 45°. But what if it is winter and the sun does not reach this angle and if traveling to another point of the earth is too expensive? Now we need more geometrical knowledge, taking into account the actual angle of the sun and the observed length of the shadow. This will yield an exact measure of the *size* of the invisible man as well.

Determining a causal effect we face the same kind of problems. First, we have to define a *causal effect*, and second, we have to find out how to determine it from empirical estimable parameters such as true means, i. e., from expectations. The simple solution — corresponding to the 45° angle of the sun in the metaphor — is the perfect randomized experiment. The sample mean differences we see in a randomized experiment only randomly deviate from the causal effect (due to random sample variation). In contrast, in quasi-experiments and observational studies, solutions to the practical problem are more sophisticated. They are also more sophisticated than in the problem of the invisible man, because it is not only *one* other variable (the angle) that determines the length of the shadow; instead there often are *many* other variables systematically determining the sample means as well as the true means that are estimated by these sample means. This is again the problem of multiple determinacy.

This book presents a solution to the theoretical and the practical problems mentioned above. Unfortunately, both solutions are not as simple and obvious as in our metaphor. Furthermore, there is not only one single kind of causal effects. (In the paragraphs above we referred to total causal effects.) To our knowledge, the first pioneer tackling the theoretical *and* the practical problems was Jerzy S. Neyman (1923/1990).

Individual and Average Causal Effects

While Fisher introduced the design technique of randomization, Neyman introduced the concepts of individual and average causal effects, thus attempting a first solution to the theoretical problem mentioned above. (Note, however, that he used different terms for these concepts). He assumed that, for each individual plot, there is an intra-individual (plot-specific) distribution of the outcome variable, say Y, under each treatment. He then simply defined the *individual causal effect of treatment x compared to treatment x'* to be the difference between the intra-individual (plot-specific) expectation of Y (the "true yield") given treatment ("variety") x and the intra-individual (plot-specific) expectation of Y given

treatment ("variety") x'. Having defined the individual causal effect, the *average treatment effect* is simply the expectation of the corresponding individual (plot-specific) causal effects in the population of observational units (plots). Similarly, several kinds of *conditional effects* can be defined, conditioning, for instance, on covariates, i. e ., on other causes of Y that cannot be affected by X, such as measures of the *quality of the soil*, *average hours of sunshine*, *average hours of rain*, etc.

Total, Direct, and Indirect Effects

At about the same time as Neyman and Fisher developed their ideas, Sewall Wright (Wright, 1918, 1921, 1923, 1934, 1960a, 1960b) developed his ideas on path analysis and the concepts of total, direct, and indirect effects. While his *total effect* aims at the same idea as the average causal effect, his *direct* and *indirect effects* were new. Simply speaking, in the context of an experiment or quasi-experiment, a direct effect of the treatment is the effect that is not transmitted through an intermediate variable; it is the conditional effect of the treatment variable holding constant the intermediate variable on one of its values. In contrast, the *indirect effect* is the difference between the total effect and the direct effect.

Fundamental Problem of Causal Inference

Whereas the basic ideas outlined above are relatively simple and straightforward, trying to put them into practice — i. e., solving the practical problem mentioned above — is often difficult and needs considerable sophistication. The "fundamental problem of causal inference" (Holland, 1986) is that we cannot expose an observational unit to treatment 1 and, at the same time, to treatment 0. However, this is exactly what is necessary if we want to be sure that 'everything else is invariant', a clause that is also an implicit idea in the solution proposed by Neyman.

Pre-Post Designs

If we choose to first observe a unit under 'no treatment' and then observe it again after 'treatment', we may be tempted to interpret the pre-post differences as estimates of the individual causal effects of the treatment given in between. However, this interpretation might be wrong, because the unit may have developed (maturated, learned), may have suffered from critical life events, may have experienced historical change, etc. (see, e. g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish et al., 2002). Hence, in these *pre-post designs* or synonymously, *within-group designs*, we have to make assumptions on the nature of these possible alternative interpretations of the pre-post comparisons, e. g., that they do not hold in the application considered or that they have a certain structure that can be taken into account when making causal inferences based on pre-post comparisons.

Between-Group Designs

If, instead of making comparisons within a unit, we compare different units to each other in *between-group experiments*, we certainly lose the possibility of estimating the *individual* causal effects. However, what we can hope for is that we are still able to estimate the *average causal effect* and certain conditional causal effects. But how to estimate the average of the individual causal effects if the individual causal effects are not estimable? Both, between-group experiments and quasi-experiments, have a set of (observational) units, at least two experimental conditions ('treatment conditions', 'expositions', 'interventions', etc.), and at least one outcome variable ('response', 'criterion', 'dependent variable') Y. In the medical sciences, the units are usually patients. In psychology the observational units are often persons, but it could be persons-in-a-situation, or groups as well. In economics it could be subjects, companies, or countries, for instance. In educational sciences the units might be school classes, schools, communities, districts, or countries. In sociology and the political sciences, the units could be persons, but also communities, countries, etc.

Scope of the Theory

In order to delineate the scope of the theory, consider the following kind of *random experiment*: Draw an observational unit u (e. g., a person) out of a set of units, observe the value z of a (possibly multivariate qualitative or quantitative) covariate Z for this unit, assign the unit or observe its assignment to one of several experimental conditions, observe the value m of an intermediate variable M, and record the numerical value y of the outcome variable Y. We will use U to denote the random variable representing with its value u the unit drawn. Note that many observations can be made additional to observing U, Z, X, M, and Y. Although this simple single-unit trial is a prototype of the kind of empirical phenomena the theory is dealing with, there are other single-unit trials in which the theory can be applied as well (see ch. 2). In fact, the theory is applicable far beyond the true experiment and the quasi-experiment. This includes applications in which the putative causes are *not* manipulable and in which the putative cause is a continuous random variable. The theory has its limitations only if there is no clear ordering of the random variables considered as putative causes or outcomes.

True Experiments and Quasi-Experiments

The single-unit trial described above is a random experiment, but not necessarily a randomized experiment. A *randomized experiment* is a special random experiment in which the unit drawn is *randomly assigned* to one of the treatment conditions, e. g., depending on the outcome of a coin toss. (In empirical applications, the single-unit trials are repeated n times, where n denotes the sample size.) Referring to single-unit trials, we can distinguish the *true experiment* from

the *quasi-experiment* as follows: In the *true experiment*, there are at least two treatment conditions and the assignment to one of the treatment conditions is randomized, e. g., by flipping a coin. In a traditional *randomized experiment*, for instance, the treatment probabilities are chosen to be equal for all units. However, equal treatment probabilities for all units are neither essential for the definition of the true experiment nor for drawing valid causal inferences. We may as well have treatment probabilities depending on the units and/or on another covariate (see section 7.5), as long as these treatment probabilities are fixed or known by the researcher. Note, however, that in designs, in which different units have different treatment probabilities, standard data analysis techniques such as t-tests or analysis of variance do not test the correct hypotheses any more.

For between-group designs, the *quasi-experiment* may be defined such that there are at least two treatment conditions; however, in contrast to the true experiment, the treatment probabilities are unknown. Nevertheless, valid causal inferences can be drawn in quasi-experiments *provided that we can rely on certain assumptions*. In specific applications these assumptions might be wrong. If they are actually wrong, causal inferences can be completely wrong as well.

Beyond Experiments and Quasi-Experiments

As it turns out, formalizing the ideas outlined above in probabilistic terms results in a theory of probabilistic causality that is applicable far beyond experiments and quasi-experiments, thus bringing together the experimental tradition of Fisher and Neyman on one side and Wright's observational studies tradition on the other side. Furthermore, causal dependencies of manifest variables measuring latent variables as well as causal dependencies between latent variables can be treated in the framework presented in this book. Hence, the scope of the theory also includes what in the past has been addressed only within structural equation modeling (see, e. g., Bentler & Wu, 2002; Jöreskog & Sörbom, 1996/2001; Muthén & Muthén, 1998-2007) and/or graphical modeling (see, e. g., Pearl, 2009; Spirtes, Glymour, & Scheines, 2000). Furthermore, specific psychometric problems such as 'differential item functioning' and 'measurement invariance' turn out to be problems of causal modeling that can be treated within the same theoretical framework as the analysis of causal effects in experimental and quasi-experimental designs.

Who Should Study This Book?

The Methodologist

In the first place, we would like to address the *methodologist*, i. e., the expert in empirical research methodology, especially in the social, economic, behavioral, cognitive, medical, agricultural, and biological sciences. This book provides answers to some of the most important and fundamental questions of these empirical sciences: What do we mean by terms like 'X affects Y', 'X has an effect on Y',

'X influences Y', 'X leads to Y', etc. used in our informal theories and hypotheses? How can we translate these terms into a language that is compatible with the statistical analysis of empirical data? How to design a study and how to look at the resulting data if we want to probe our theories empirically and learn about the causal dependencies postulated in these theories and hypotheses? And last but not least: How to evaluate interventions, treatments, or expositions to (possibly detrimental) environments and learn about how effective they are for which kind of subjects or observational-units, and under which circumstances?

The Statistician

Many statisticians believe that causality is beyond their horizon. Causality might be a matter of empirical researchers and philosophers, they say, but not their own. They think that it cannot be treated mathematically and therefore a statistician cannot be helpful. As a consequence, they ignore the issue of causality. Reading this book will prove that all these beliefs should be abandoned. Probabilistic causality, as presented here, is a branch of probability theory, which itself, at least since Kolmogorov (1956), is a part of pure mathematics — although with an enormous potential for applications in many empirical sciences and even beyond. The main purpose of this book is to translate the informal concepts about causality shared by many methodologists and applied statisticians into the well-defined terms of mathematical probability theory. The principle is not to use any undefined term, and the result is a pure mathematical theory of probabilistic causality. Of course, this will make it harder for the methodologist and those not yet trained in probability theory. However, the reward is a much deeper understanding of what is essential and a much better grasp of the nature of our theories about the real world.

Of course, undefined terms are still used in this book, but only in the examples, in the interpretations, and in the motivations of the definitions. The theory itself is pure mathematics, just in the same way as Kolmogorov's probability theory presented in 1933, which explicated the mathematical, measure-theoretical structure of probabilistic concepts. Substantive meaning only results if we interpret the core components of the formal structure in a specific random experiment considered. And this is also true for the theory of probabilistic causality presented in this book.

The Empirical Scientist

The empirical scientist in the fields mentioned above has at least two good reasons to study this book. The *first* is that some crucial parts of his theories and hypotheses are explicated, at least when it comes to considering a concrete experiment or study. The ambiguity in causal language such as 'X affects Y', 'X has an effect on Y', 'X influences Y', 'X leads to Y' are not necessary any more. Reading this book will make it possible to replace these ambiguous terms by well-

understood and well-defined terms, improving quality of empirical research and theories.

The *second* motivation of the empirical scientist is that even if he knows his own theoretical concepts and hypotheses, he still has to know how to design experiments and studies that enable him to test them. Furthermore, the standard ways of analyzing data offered in the textbooks of applied statistics and in the available computer programs often do not estimate and test the correct causal effects and dependencies. And this is not only bad for the empirical scientist but also for all those relying on the validity of his inferences and his expertise. Just think about all the harmful consequences of wrong causal theories in various empirical research fields, if they are applied to solving concrete problems!

The Experimental Scientist

This book has two messages for those who do their research with experiments, a good one and a bad one. The good news is that, in perfect randomized experiments, the average causal total treatment effect is indeed estimated when comparing means between two different treatment conditions. The bad news is that *we can not rely on randomized assignment of units to treatment conditions* when it comes to estimating *direct* and *indirect* effects. More specifically, in such an analysis it is usually not sufficient to consider intermediates, treatment and outcome variables. Instead we also have to include in our analysis *pre-treatment variables* such as a pre-test of the intermediate and a pre-test of the outcome variable and apply adjustment methods, very much in the same way as we have to use these techniques in quasi-experiments — *even though we have randomized*! Hence if you want to look into the black box between the treatment and the outcome variables, you have to adopt the techniques of causal modeling that are far beyond traditional comparisons of means and analysis of variance.

The Philosopher of Science

Philosophers of science study and teach the methodology of empirical sciences. In that respect, their task is very similar to that of the methodologist, perhaps only more general and less specific for a certain discipline. Therefore, it is not surprising that probabilistic causality has also been tackled by philosophers of science (see, e. g., Cartwright, 1979; Spohn, 1980; Stegmüller, 1983; Suppes, 1970). Compared to these approaches, our emphasis is more on those parts of the theory that have implications for the *design* of empirical studies and the *analysis of data* resulting from such studies.

The Students in These Fields

We believe that probabilistic causality is the most rewarding topic in methodology. Although it is tough to get into it, you will get insights why all this methodology stuff was useful and what it was good for. At least this is what our students

say at the end of our curriculum, even if they did not have the choice whether or not to take our course on probabilistic causality.

Research Traditions in Stochastic Causality

Several research traditions have been contributing to the theory probabilistic causality in various ways. From the *Neyman-Rubin tradition*, we adopted the idea that it is important to define various causal effects such as individual, conditional, and average causal effects, even though we modified and extended these concepts in important aspects. Defining causal effects is important for proving that certain methods of data analysis yield estimates of these effects if certain assumptions can be made. Are there conditions under which the analysis of change scores (between pre- and post-tests) and repeated-measures analysis of variance yield causal effects? Under which conditions do we test causal effects in the analysis of covariance? Which are the assumptions under which propensity score methods yield estimates of causal effects? Which are the assumptions under which an instrumental variable analysis estimates a causal effect? All these questions and their answers presuppose that we have a clear definition of causal effects and/or of causal probabilistic dependencies.

From the *Campbellian tradition* (see, e. g., Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002) we learned that there are questions and problems beyond stochastic causality itself that are relevant in empirical causal research, such as: How to generalize beyond the study? What does the treatment variable mean? What is the meaning of the outcome variable? And, perhaps the most important question: Are there alternative explanations for the effect? The vast majority of social scientists (including ourselves) have been educated in this research tradition to some degree. Although this training is still very useful as a general methodology framework, it lacks precision and clarity in a number of issues — and causality is one of these.

From the *graphical modeling tradition* (see, e. g., Cox & Wermuth, 2004; Pearl, 2009; Spirtes et al., 2000), we learned that conditional independence plays an important role in causal modeling. This research tradition has also been developing techniques to estimate causal effects and to search for causal models if specific assumptions can be made. The fact that randomization in a true experiment in no way guarantees the validity of causal inferences on *direct* effects has been brought up by this research tradition.

Structural equation modeling and *psychometrics* have been teaching us how to use latent variables and structural equation modeling in testing causal hypotheses. Due to a number of statistical programs such as AMOS (Arbuckle, 2006), EQS (Bentler, 1995), lavaan (?, ?), LISREL (Jöreskog & Sörbom, 1996/2001), Mplus (Muthén & Muthén, 1998-2007), OpenMx (OpenMx, 2009), RAMONA (Browne & Mels, 1998), structural equation modeling became extremely popular in the Social Sciences. Although many users of these programs hope to find causal answers, it should be clearly stated that structural equation modeling — and this is true for all kinds of statistical models (including analysis of variance) — does

neither automatically estimate and test causal effects, nor does it provide a satisfactory *theory* of causal effects and dependencies. Nevertheless, this research tradition contributes — just like other areas of statistics — a number of statistical techniques that can be very useful in causal modeling.

In this book, we also aim at embedding — and, where necessary, extending — conventional statistical procedures such as analysis of covariance, nonorthogonal analysis of variance, and latent variable modeling, but also more recent techniques based on propensity scores, or on instrumental variables into a coherent theory of probabilistic causality.

How to Use This Book

This book is self-contained. It is written such that standard mathematical probability theory is sufficient for a complete understanding, provided one takes the time that these topics require. In many parts, this is not a book one can just *read*; instead it is a book to be *studied*. This includes working on the questions and exercises. We presume that the reader is familiar with — or learns while studying this book — the essentials of probability theory, including conditional expectations, as well as conditional independence and conditional distributions. These essentials of probability theory are dealt with in Steyer and Nagel (in press-a).

We devoted this book almost entirely to the *theory* of causal effects and probabilistic causality, although, in chapter 13, we outline the implications of the theory for *design* and for *data analysis in experiments and quasi-experiments*. We also developed the PC program *Causal Effects Explorer* (Nagengast, Kröhne, Bauer, & Steyer, 2007) that can be used for exploring prima facie effects, conditional and average total effects given certain parameters. We believe that this program is useful for teaching and learning the fundamentals of the theory. Furthermore, the program *EffectLiteR* (?, ?), can be used to estimate total, direct, and indirect effects from empirical data in experiments and quasi-experiments. Both programs, which are available at *www.causal-effects.de*, may be used together with this book in a course on causal modeling. In fact, this is the content of our workshops on the analysis of total, direct, and indirect causal effects, which are available both as videos-on-demand on the internet and on DVDs, again at *www.causal-effects.de*.

Acknowledgements

This book has been written with the help of several colleagues and students. Werner Nagel (FSU Jena) helped whenever we felt lost in probability spaces. Stephen G. West (Arizona State University) and Felix Thömmes (University of Tübingen) made detailed suggestions for improving readability of the book. Safir Yousfi and Sonja Hahn contributed and/or suggested concrete ideas, Sonja being extremely helpful also in checking some of the mathematics. Our students Lisa Dietzfelbinger, Niclas Heider, Marc Heigener, Remo Kamm, Lawrence Lo, Axel Mayer, David Meder, Marita Menzel, Yuka Morikawa, Sebastian Nitsche, Michael

Temmerman, Sebastian Weirich, Anna Zimmermann, and other students critically commented on previous versions, helped minimizing errors, or organizing the references. Sven Hartenstein, Christoph Nachtigall, Marc Müller, Steffi Pohl, Norman Rose and Andreas Wolf together with the others mentioned above provided the intellectual climate in which this book could be written. We are also grateful to the students and colleagues participating at our courses on the analysis of causal effects asking questions and making important comments. Over the years, this helped a lot to improve this book.

Jena, April 15, 2016

Contents

List of Figures

List of Tables

Chapter 1
Introductory Examples

For more than a century there have been examples in the statistical literature showing that comparing means or comparing probabilities (e. g., of success of a treatment) between a group exposed to a treatment and a comparison group (unexposed or exposed to a different treatment) does not necessarily answer our questions: 'Which treatment is better overall?' or 'Which treatment is better for which kind of person?' Differences between means and differences between probabilities (or any other comparison between probabilities such as odds ratios, log odds ratios, or relative risk) are usually not the treatment effects we are looking for (see, e. g., Pearson, Lee, & Bramley-Moore, 1899; Yule, 1903; Simpson, 1951). They are just *effects at first sight* or "prima facie effects" (Holland, 1986).

Just like the shadow in the metaphor of the invisible man (see the preface), prima facie effects reflect the effects of the treatment (the size of the invisible man), but also of other causes (the angle of the sun). The goal of analyzing *causal* effects is to estimate the effect of the treatment alone, isolating it from other potential influences, e. g., of sex, educational background, socio-economic status, etc. The general idea is to compute a treatment effect that is not biased by differences between treatment groups that would also exist *without treatment*.

Overview

We will illustrate systematic bias in determining *total* treatment effects in quasi-experiments by two examples. The first one deals with a dichotomous outcome variable, the second with a quantitative one. While the problems described in these two examples cannot occur in a randomized experiment, our third example will show that the randomized assignment of units to treatment conditions does not help to prevent systematic bias in determining *direct* treatment effects with respect to an intermediate variable that may transmit the effects of the treatment on the outcome variable.

1.1 Example 1 — Simpson's Paradox

In our first example, the prima facie effect reverses if we switch from comparing $P(Y{=}1\,|X{=}1)$ to $P(Y{=}1\,|X{=}0)$, the conditional probabilities of success between treatment and control, to comparing $P(Y{=}1\,|X{=}1, Z{=}z)$ to $P(Y{=}1\,|X{=}0, Z{=}z)$,

Table 1.1. Joint Probabilities of Treatment and Success

	Treatment		
Success	No (X=0)	Yes (X=1)	
No (Y=0)	.240	.232	.472
Yes (Y=1)	.360	.168	.528
	.600	.400	1.000

the corresponding probabilities additionally controlling for Z = *sex* with values m (males) and f (females). This kind of phenomenon, which is already known at least since Yule (1903), is called *Simpson's paradox* (Simpson, 1951), and it is still being debated (see, e. g., Hernán, Clayton, & Keiding, 2011).

1.1.1 Prima Facie Effect

Table 1.1 shows the joint distribution of treatment and success, i. e., the joint probabilities $P(X{=}x, Y{=}y)$ of treatment and success, as well as the marginal probabilities $P(X{=}x)$ and $P(Y{=}y)$ of treatment x and success y, respectively. Comparing the conditional probability of success (Y=1) given the *treatment condition* (X=1) to the conditional probability of success given the *control condition* (X=0) would lead us to the conclusion that the *treatment is harmful*. These two conditional probabilities can be computed by

$$P(Y{=}1\,|\,X{=}1) = \frac{P(Y{=}1, X{=}1)}{P(X{=}1)} = \frac{.168}{.168 + .232} = .42$$

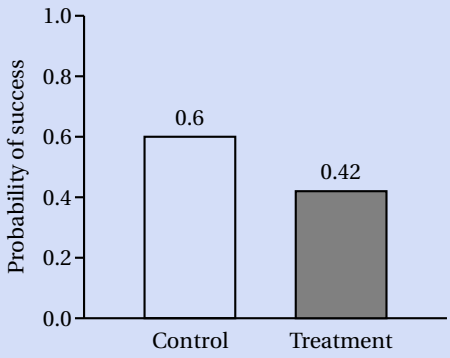

Figure 1.1. Probability of success given treatment

and

$$P(Y=1\,|\,X=0) \;=\; \frac{P(Y=1,X=0)}{P(X=0)} \;=\; \frac{.360}{.360+.240} \;=\; .60,$$

respectively (see, e. g., Steyer & Nagel, in press-a, section 4.2). Figure 1.1 displays both conditional probabilities in a histogram.

These two conditional probabilities can be compared to each other in different ways. The simplest one is looking at the *difference* $P(Y=1\,|\,X=1)-P(Y=1\,|\,X=0)$. This is a particular case of the difference $E(Y\,|\,X=1)-E(Y\,|\,X=0)$ between two conditional expectation values, in which the outcome variable Y is dichotomous with values 0 and 1. Following Holland (1986), we will call this difference the (unconditional) *prima facie effect* and use the notation PFE_{10}. Other possibilities of comparing the two conditional probabilities are to look at the odds ratio, or the logarithm of the odds ratio (see chapter 4 of Rothman, Greenland, & Lash, 2008, for a detailed discussion of these and other effect parameters).

1.1.2 Prima Facie Effects Controlling for Sex

The conclusion about the effect of the treatment is completely different if we look at the dependencies separately for males and females. Table 1.2 (p. 4) shows the joint distributions of treatment, success and $Z :=$ *sex* with values 0 (*male*) and 1 (*female*). The probabilities of the two values are $P(Z=0) = P(Z=1) = .50$. According to this table, the probability of success for the males in the treatment condition is

$$P(Y=1\,|\,X=1,Z=0) \;=\; \frac{.016}{.016+.004} \;=\; .80$$

(see Exercise 1-7), whereas the probability of success in the control condition is

$$P(Y=1\,|\,X=0,Z=0) \;=\; \frac{.336}{.336+.144} \;=\; .70.$$

Hence, the difference

$$P(Y=1\,|\,X=1,Z=0) \;-\; P(Y=1\,|\,X=0,Z=0) \tag{1.1}$$

is $.80-.70 = .10$, which may lead us to conclude that *the treatment is beneficial for males*. Again, because Y is dichotomous with values 0 and 1, this difference is a particular case of the difference $PFE_{10;Z=0} := E(Y\,|\,X=1,Z=0)-E(Y\,|\,X=0,Z=0)$, which we call the *conditional prima facie effect* given $Z=0$.

What about the treatment effects for females? Table 1.2 shows that the probability of success for the females in the treatment condition is $.152/(.152+.228) = .40$, whereas it is $.024/(.024+.096) = .20$ in the control condition. Figure 1.2 shows these conditional probabilities in a histogram. Considering the difference $.40-.20 = .20$ may lead us to conclude that *the treatment is also beneficial for females*.

Hence, we can conclude that the treatment seems to be *beneficial for both, males and females*. This, however, seems to contradict our finding ignoring sex. Just considering the difference $E(Y\,|\,X=1)-E(Y\,|\,X=0)$, the *treatment seemed to be harmful*.

Table 1.2. Joint Probabilities of Treatment, Sex and Success

Males ($Z=0$); $P(Z=0)=0.50$

	Treatment		
Success	No ($X=0$)	Yes ($X=1$)	
No ($Y=0$)	.144	.004	.148
Yes ($Y=1$)	.336	.016	.352
	.480	.020	.500

Females ($Z=1$); $P(Z=1)=0.50$

	Treatment		
Success	No ($X=0$)	Yes ($X=1$)	
No ($Y=0$)	.096	.228	.324
Yes ($Y=1$)	.024	.152	.176
	.120	.380	.500

1.1.3 Prima Facie Effect vs. Average of the Prima Facie Effects

In contrast to our intuition, the *prima facie effect* $E(Y|X=1)-E(Y|X=0)$ is *neither* the *simple average nor any weighted average of the corresponding prima fa-*

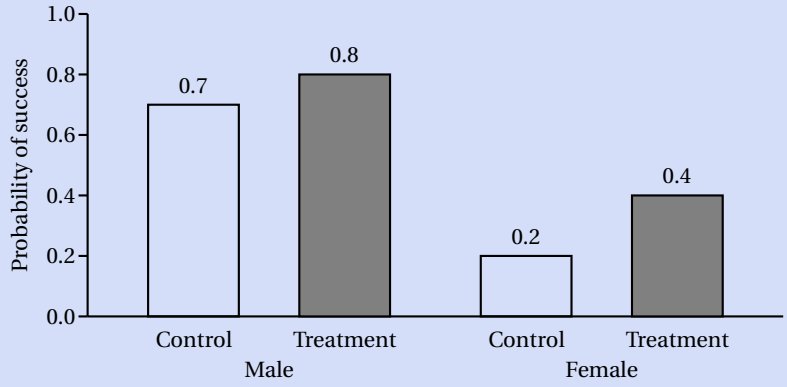

Figure 1.2. Probabilities of success given treatment (and sex)

cie effects $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$ controlling for Z = *sex*. This is now studied in more detail.

Prima Facie Effect

The probability $P(Y=1|X=0)$ of success in the control condition is the sum of the corresponding probabilities, $P(Y=1|X=0,Z=0)$ and $P(Y=1|X=0,Z=1)$, *weighted by the conditional probabilities* $P(Z=0|X=0)$ and $P(Z=1|X=0)$, respectively, i. e.,

$$\begin{aligned} P(Y=1\,|\,X=0) \;&= P(Y=1\,|\,X=0,Z=0)\cdot P(Z=0\,|\,X=0) + \\ &\quad\; P(Y=1\,|\,X=0,Z=1)\cdot P(Z=1\,|\,X=0) \\ &= .70\cdot\frac{.48}{.60} \;+\; .20\cdot\frac{.12}{.60} \;=\; .60 \end{aligned}$$

[see Box 9.2 (ii) of Steyer & Nagel, in press-a, and Exercise 1-8]. Because the difference between the conditional probabilities $P(Z=0\,|\,X=0)$ = .48/.60 and $P(Z=1\,|\,X=0)$ = .12/.60 is large, the probability of success in treatment 0 is much closer to .70 than to .20 (see the dots above $X=0$ in Fig. 1.3).

Similarly, the probability $P(Y=1\,|\,X=1)$ of success in the treatment condition ($X=1$) is the sum of the two corresponding probabilities, $P(Y=1\,|\,X=1,Z=0)$ and $P(Y=1\,|\,X=1,Z=1)$, *weighted by the conditional probabilities* $P(Z=0\,|\,X=1)$ and $P(Z=1\,|\,X=1)$, respectively, i. e.,

$$\begin{aligned} P(Y=1\,|\,X=1) \;&= P(Y=1\,|\,X=1,Z=0)\cdot P(Z=0\,|\,X=1) + \\ &\quad\; P(Y=1\,|\,X=1,Z=1)\cdot P(Z=1\,|\,X=1) \\ &= .80\cdot\frac{.02}{.40} + .40\cdot\frac{.38}{.40} \;=\; .42. \end{aligned}$$

Hence, the *prima facie effect* is $P(Y=1\,|\,X=1)-P(Y=1\,|\,X=0)$ = .42 − .60. = −.18. Because the two conditional probabilities $P(Z=0\,|\,X=1)$ = .02/.40 and $P(Z=1\,|\,X=1)$ = .38/.40 are very different, the probability of success in treatment 1 is much closer to .40 than to .80 (see the dots above $X=1$ in Fig. 1.3). (The size of the area of the dotted circles represent the joint probabilities $P(X=x,Z=z)$. For $X=1$ and $Z=0$, this probability is very small such that the circle is not visible. This kind of graphics has been adopted from Agresti, 2007).

Average of the Conditional Prima Facie Effects

In contrast to the prima facie effect, the *average of the conditional prima facie effects* is the expectation of the function $PFE_{10;Z}$, the values of which are the two prima facie effects $PFE_{10;Z=0}$ and $PFE_{10;Z=1}$ for males and females, i. e.,

$$E(PFE_{10;Z}) \;=\; \sum_z PFE_{10;Z=z}\cdot P(Z=z). \tag{1.2}$$

Because the conditional prima facie effect of the treatment is $PFE_{10;Z=0}$ = .10 for males and $PFE_{10;Z=1}$ = .20 for females, the average prima facie effect is simply:

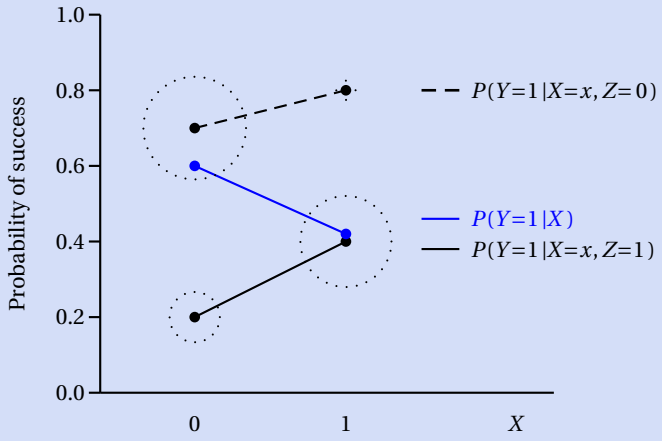

Figure 1.3. Probabilities of success given treatment and sex

$$E(PFE_{10;Z}) \;=\; .10\cdot P(Z=0) + .20\cdot P(Z=1) \;=\; .10\cdot\frac{1}{2} + .20\cdot\frac{1}{2} \;=\; .15\,.$$

Hence, whereas the *prima facie effect* $E(Y|X=1)-E(Y|X=0)$ is *negative*, namely $-.18$, the *average of the* $(Z=z)$*-conditional prima facie effects* $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$ is *positive*, namely .15.

1.1.4 How to Evaluate the Treatment?

Because the conclusions drawn from the differences $E(Y|X=1)-E(Y|X=0)$ and $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$ are contradictory, which of these comparisons should we trust? Is the treatment harmful — as $E(Y|X=1)-E(Y|X=0)$ suggests? Or is it beneficial as suggested by the differences $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$? Which of these comparisons are meaningful for evaluating the causal effect of the treatment? Before we come back to these questions, let us consider another example.

1.2 Example 2 — Nonorthogonal Two-Factorial Experiment

In this section, we treat an example with three treatment conditions, three values of a discrete covariate, and a quantitative outcome variable. In this example, we use a 3×3 factorial design with crossed, non-orthogonal factors. The analysis of such designs has been puzzling many statisticians (see, e. g., Aitkin, 1978; Appelbaum & Cramer, 1974; Carlson & Timm, 1974; Gosslee & Lucas, 1965; Jennings & Green, 1984; Keren & Lewis, 1976; Kramer, 1955; Overall & Spiegel, 1969, 1973b,

Table 1.3. Conditional Expectations Given Treatment

Treatment	Expectation of Y in treatment conditions $E(Y\|X{=}x)$	Treatment probabilities $P(X{=}x)$
$X{=}0$ (Control)	111.25	1/3
$X{=}1$ (Treatment 1)	100.00	1/3
$X{=}2$ (Treatment 2)	114.25	1/3
$E(Y)$	108.50	

1973a; Overall, Spiegel, & Cohen, 1975; Williams, 1972), and it continues to do so (see, e. g., Langsrud, 2003; Nelder & Lane, 1995).[1]

1.2.1 Prima Facie Effects

In the example presented in Table 1.3, there are *three treatment conditions* representing two treatments and a control. The outcome variable Y is now a quantitative measure of success. The expectations of the outcome variable Y in the three treatment conditions are displayed in Table 1.3. The ratios in the last column are the treatment probabilities $P(X{=}x)$ which are, in this example, the same for all three treatment conditions. However, although the probabilities $P(X{=}x)$ are the same for all three groups, this is *not* a randomized design as will become obvious if we look at the second factor and the 'cell probabilities' (see Table 1.4). Discussing the example at the level of conditional expectation values will again make clear that the contradictory inferences are not due to errors in *statistical inference* (from sample statistics to true parameters), but due to errors in *causal inference*, i. e., they are due to the misinterpretation of the differences between the expectations $E(Y|X{=}x)$ of the outcome variable Y in the three treatment conditions as causal effects.

If our evaluation of the treatment effects were based on these differences between the expectations of Y in the three treatment conditions, we would conclude that there are two treatment effects: a *negative effect* (namely, $100.00 - 111.25 = -11.25$) of treatment 1 compared to the control, and a *positive effect* (namely, $114.25 - 111.25 = 3.00$) of treatment 2 compared to the control.

[1] In fact, none of the statistical packages such as SAS, SysStat, or SPSS with their Type I, II, III or IV sums of squares provide correct estimates and tests of the average effects (or main effects) for such a design unless the covariate (the second factor) has a uniform distribution, with equal probabilities for all values of the covariate. In this case Type III analysis yields correct results, at least, if the second factor is assumed to be fixed. However, in most applications in the Social Sciences, the covariate (second factor) is not fixed but stochastic with varying sample means, etc. In chapter 13, we will outline a correct analysis including the average total effects.

Table 1.4. Conditional Expectations Given Treatment and Neediness

	Neediness						
Treatment	Low (Z=0)		Medium (Z=1)		High (Z=2)		
X=0	120	(20/120)	110	(17/120)	60	(3/120)	(40/120)
X=1	100	(7/120)	100	(26/120)	100	(7/120)	(40/120)
X=2	80	(3/120)	90	(17/120)	140	(20/120)	(40/120)
		(30/120)		(60/120)		(30/120)	

Note. Probabilities $P(X{=}x, Z{=}z)$, $P(Z{=}z)$, and $P(X{=}x)$ in parentheses.

1.2.2 Prima Facie Effects Controlling for Neediness

A second way to evaluate the 'effects' of the three *treatment conditions* is to look at the differences between the expectations of Y in the three treatment conditions *within each of the three classes of neediness* for the therapy: low, medium, and high. Table 1.4 displays the expectations of the outcome variable Y in the nine cells of the 3×3 design. The ratios in parentheses are the probabilities that the pairs (x, z) of values of X and Z are observed. Hence, this table contains the conditional expectation values (true cell means) of the outcome variable Y, and the probabilities $P(X{=}x, Z{=}z)$ determining the true joint distribution of X and Z.

In the *low neediness condition* ($Z{=}0$), there are large negative effects, both of treatment 1 and of treatment 2 compared to the control:

$$PFE_{10;Z=0} := E(Y|X{=}1, Z{=}0) - E(Y|X{=}0, Z{=}0) = 100 - 120 = -20$$

and

$$PFE_{20;Z=0} := E(Y|X{=}2, Z{=}0) - E(Y|X{=}0, Z{=}0) = 80 - 120 = -40.$$

In the *medium neediness condition* ($Z{=}1$), there are also negative effects of treatment 1 and of treatment 2 compared to the control:

$$PFE_{10;Z=1} := E(Y|X{=}1, Z{=}1) - E(Y|X{=}0, Z{=}1) = 100 - 110 = -10$$

and

$$PFE_{20;Z=1} := E(Y|X{=}2, Z{=}1) - E(Y|X{=}0, Z{=}1) = 90 - 110 = -20.$$

Finally, in the *high neediness condition* ($Z{=}2$), the effects of treatment 1 and treatment 2 are both positive:

$$PFE_{10;Z=2} := E(Y|X{=}1, Z{=}2) - E(Y|X{=}0, Z{=}2) = 100 - 60 = 40$$

and

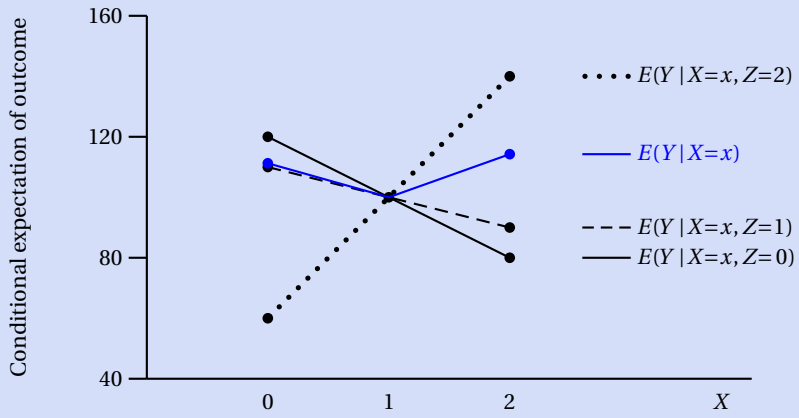

Figure 1.4. Conditional expectation values of Y given treatment and neediness

$$PFE_{20;Z=2} := E(Y|X=2,Z=2) - E(Y|X=0,Z=2) = 140 - 60 = 80.$$

Based on these comparisons, we can conclude that the 'effects' of the treatments depend on the neediness of the subjects: the differences between the expectations of Y are negative for subjects with low and medium neediness, and they are positive for the subjects with high neediness.

1.2.3 Prima Facie Effects vs. Average of the Prima Facie Effects

There is no doubt that the conditional effects given neediness, which are sometimes also called *simple effects*, are more informative than average treatment effects if we want to know which treatment is the best for which level of neediness. Nevertheless, we might ask: What are the 'treatment effects' on average? Or, in other words which are the 'main effects'? In fact, all major statistical programs compute 'main effects' (see Langsrud, 2003 for a list on which program suggests what solution to this problem). Note that we have two average effects in this example, because we can compare treatment 1 *and* treatment 2 to the control. Because we already looked at the corresponding conditional effects, we just have to compute their averages, i. e., the expectations of these conditional effects over the distribution of neediness:

$$E(PFE_{10;Z}) = \sum_z PFE_{10;Z=z} \cdot P(Z=z) = -20 \cdot \frac{1}{4} + (-10) \cdot \frac{1}{2} + 40 \cdot \frac{1}{4} = 0.$$

Hence, the average effect of treatment 1 compared to the control is zero.

Comparing treatment 2 to the control yields on average:

$$E(PFE_{20;Z}) = \sum_z PFE_{20;Z=z} \cdot P(Z{=}z) = -40 \cdot \frac{1}{4} + (-20) \cdot \frac{1}{2} + 80 \cdot \frac{1}{4} = 0.$$

According to this result, the average effect of treatment 2 compared to the control is zero as well.

1.2.4 How to Evaluate the Treatment?

To summarize, we discussed three ways that may, at first sight, be used to evaluate the treatment effects: *First*, we may compare the differences between the expectations $E(Y|X{=}x)$ of the outcome variable in the three treatment conditions $X{=}0$, $X{=}1$, and $X{=}2$. *Second*, we may consider the corresponding differences between the conditional expectations $E(Y|X{=}x, Z{=}z)$ *within each of the three values* $Z{=}0$, $Z{=}1$, *and* $Z{=}2$ *of neediness. Third*, we may compare the averages of these differences between the conditional expectations over the distribution of Z (see Box 1.1 for a summary of these effects).[2] All these comparisons yield different results. Which of them are meaningful for the evaluation of the treatment effects? All three of them, or only two, just one, or none at all?

1.3 Example 3 — Direct Effect in a Randomized Experiment

The problems described in the examples treated in the preceding sections occur because there are covariates (in the examples, *sex* and *neediness*) that are related to the treatment variable *and* the outcome variable. Hence, these problems can *not* occur in a randomized experiment, in which, by definition, all covariates and the treatment are (stochastically) independent. Hence, if in a randomized experiment, we are only interested in the *total effects* of the treatment on the outcome variable, the effects that are estimated by the differences between means in the treatment groups are the total effects of the treatment. However, often we are also interested in the mediation processes producing these total effects. A typical question in educational research is: 'Is there a direct effect of the treatment that is not transmitted through *motivation after treatment*?' In medical research we may ask: 'Is there a direct effect of the treatment that is not transmitted through the *amount of antibodies*?'

1.3.1 Conditional Expectation of Y Given Treatment and Intermediate Variables

Suppose that Table 1.5 displays the true means, variances, covariances, and correlations of a treatment variable X with values 0 and 1, an intermediate vari-

[2] In fact, there are even more than three ways. Types II and III of computing the sums of squares in nonorthogonal ANOVA are not yet considered in our discussion. In chapter 13, we show that all four types of computing sums of squares in such a design yield wrong results in our example (see also Exercise 1-14).

Table 1.5. Covariances, Correlations, and Expectations (Omitting Pre-Tests)

		X	M	Y
Treatment (yes = 1, no = 0)	X	0.25	*.727*	*.597*
Post-test motivation	M	5.00	189.00	*.893*
Post-test achievement	Y	5.00	205.70	280.45
Expectations		0.50	90.00	140.00

Note. Correlations (in italics) are rounded.

able M, and an outcome variable Y. (This example is adopted from Mayer, Thoemmes, Rose, Steyer, & West, 2014.)

First of all, let us consider the conditional expectation $E(Y \mid X, M)$, assuming that it can be written as a linear function of X and M (see Fig. 1.5). In fact, the covariance matrix presented in Table 1.5 has been constructed such that this linearity assumption holds. Using the covariances and expectations displayed in this table, we receive

$$E(Y \mid X, M) \approx 34.9924 - 3.7528 \cdot X + 1.1876 \cdot M \tag{1.3}$$

(see Exercise 1-12). According to textbook wisdom (see, e. g., MacKinnon, 2008, but also Baron & Kenny, 1986), the direct effect of X on Y, controlling for M, is approximately -3.75.

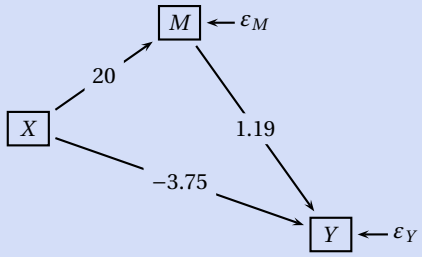

Figure 1.5. Path diagram of $E(M \mid X)$ and $E(Y \mid X, M)$

Table 1.6. Covariances, Correlations, and Expectations (Including Pre-Tests)

		W	Z	X	M	Y
Pre-test achievement	W	100.00	*.850*	*.000*	*.495*	*.740*
Pre-test motivation	Z	85.00	100.00	*.000*	*.582*	*.696*
Treatment (yes=1, no=0)	X	0.00	0.00	0.25	*.727*	*.597*
Post-test motivation	M	68.00	80.00	5.00	189.00	*.893*
Post-test achievement	Y	124.00	116.50	5.00	205.70	280.45
Expectations		100.00	100.00	0.50	90.00	140.00

Note. Correlations (in italics) are rounded.

1.3.2 Conditional Expectation of Y Given Treatment, Intermediate, and Pre-Test Variables

Suppose M represents *post-test motivation* in a randomized experiment designed to evaluate two teaching methods represented by $(X=0)$ and $(X=1)$, respectively. In this case, even if not observed, there will be a variable, say Z representing *pre-test motivation* with respect to which students will differ before treatment. Furthermore, there will be a variable, say W, representing *pre-test achievement* with respect to which students will differ prior to treatment as well. Furthermore, the two pre-test variables Z and W will be correlated. This is a plausible scenario for such a teaching experiment, and this is how the complete variance-covariance matrix and the expectations presented in Table 1.6 have been generated.

Hence, if instead of $E(Y|X,M)$, we consider the conditional expectation of Y given X, M, Z, and W, again assuming linearity — and this is how the parameters presented in Table 1.6 have been generated — we receive

$$E(Y|X,M,Z,W) = .00 + 10\cdot X + 0.50\cdot M + 0.00\cdot Z + .90\cdot W \tag{1.4}$$

(see Exercise 1-13). Now the coefficient 10 of X might be interpreted to be the direct treatment effect, 'direct' with respect to the intermediate variable M. It is the effect of X controlling for the intermediate variable M *and* for all covariates, in this example, the two pre-test variables Z and W.

How can we explain this seemingly paradoxical result? How can there be confounding in a perfect randomized experiment? The answer is that even though X and the bivariate random variable (W,Z) are independent, *conditional independence* of X and (W,Z) given M does *not* hold. Instead, conditioning on M induces conditional *dependence* of X and Z, if both Z and X are related to M. Intuitively speaking, because both Z and X affect M, a high value of *post-test motivation* M means that both, X and Z tend to be high, whereas a low value of M means that both, X and Z tend to be low (see Fig. 1.6). Hence, conditioning on M, the treatment variable X and the *pre-test motivation* Z will be dependent, even though X and Z are unconditionally independent, due to randomization (see also Pearl,

2009, ch. 1, p. 17, or Spirtes et al., 2000). This conditional dependence between X and Z given M is also reflected by a non-zero partial correlation $Corr(X, Z; M)$ (see section 11.6 of Steyer & Nagel, in press-a).

1.3.3 Conditional Expectation of Y Given Treatment Variable

Finally, let us consider the *average total treatment effect*. In this example, in which X and all covariates are independent, the average total treatment effect is the coefficient of X in the equation

$$E(Y|X) = 130 + 20 \cdot X, \tag{1.5}$$

where the intercept $\alpha_0 = 130$ is obtained by $\alpha_0 = E(Y) - \alpha_1 \cdot E(X) = 140 - 20 \cdot 0.50 = 130$ and the slope by $\alpha_1 = Cov(X, Y)/Var(X) = 5.00/0.25 = 20$ [see Steyer & Nagel, in press-a, Eqs. (12.58) and (12.59)]. Therefore, in this example, the *indirect treatment effect* is the difference $20 - 10 = 10$. In this model with no interaction, this indirect effect is also equal to the product $20 \cdot .50$ (see Fig. 1.6), which is in accordance with the rules of path analysis developed by Sewall Wright in the twenties of last century (see, e. g., Wright, 1918, 1921, 1923).

1.3.4 How to Analyze Direct Effects?

We discussed two different ways to analyze the direct effect of the treatment variable on the outcome variable. The first one is recommended in traditional textbooks such as MacKinnon (2008) and in one of the most frequently cited papers Baron and Kenny (1986). It yields the negative direct effect of −3.75. The second one also controls for the pre-tests of the intermediate variable and the outcome variables. This second analysis yields a direct treatment effect of 10. Hence, the effect is reversed as compared to the first analysis. Which is the correct direct effect? Or are both wrong?

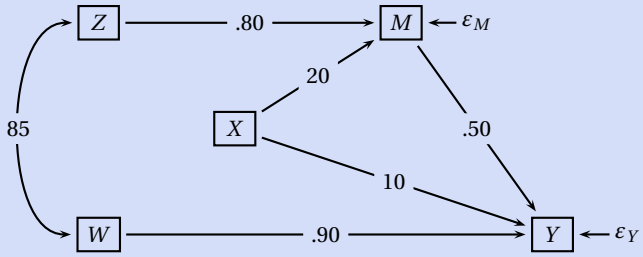

Figure 1.6. Path diagram of $E(M|X, Z, W)$ and $E(Y|X, M, Z, W)$

1.4 Summary and Conclusions

In this chapter, we treated three examples. In the first example, a dichotomous treatment variable X has a negative 'effect' $E(Y|X=1)-E(Y|X=0)$ on a *dichotomous outcome variable* Y ('success'), although the corresponding treatment 'effects' $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$ are positive if we condition on males ($Z=m$) and females ($Z=f$). Taking the expectation of these two conditional effects also yielded a positive 'effect'. In the second example, there are nonzero differences $E(Y|X=1) - E(Y|X=0)$ and $E(Y|X=2) - E(Y|X=0)$, where Y is a *quantitative outcome variable*, and nonzero conditional 'effects' $E(Y|X=1,Z=z) - E(Y|X=0,Z=z)$ and $E(Y|X=2,Z=z) - E(Y|X=0,Z=z)$ for the different values of *neediness*. The expectations of these conditional 'effects' over the three neediness conditions, i. e., the average 'effects', are zero. In the third example, we discussed two different ways of analyzing the direct treatment effect. The first yields a negative 'direct effect' and the second a positive 'direct effect'.

The Problem

Because the conclusions drawn from these analyses are contradictory, which of these should we trust? In Simpson's paradox: Is the treatment harmful — as the difference $E(Y|X=1)-E(Y|X=0)$ suggests? Or is it beneficial as suggested by the differences $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$, controlling for sex? Which of these comparisons are meaningful for the evaluation of the causal effects of the treatment? Similarly, in the second example: are there treatment effects, overall? Or are the effects nil on average? And, are the conditional effects dependable, or could it be that they would also be reversed if we condition on an additional covariate, such as *age* or *educational status*? As demonstrated in Simpson's paradox, we can neither expect that the difference $E(Y|X=1)-E(Y|X=0)$ is the average of the corresponding differences $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$, nor can we expect that a difference $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$ is the average over the corresponding differences if we condition on an additional covariate such as age. Note, these questions are not related to *statistical* inference; they are not raised at the sample level, but on the level of true parameters!

Hence our examples show that the conditional expectation values and their differences, the prima facie effects, can be totally misleading in evaluating the effects of a treatment variable X on an outcome variable Y. This conclusion can also be extended to conditional probabilities, to correlations and to all other parameters describing relationships and dependencies between random variables. They all are like the shadow in the metaphor of the invisible man (see the preface).

If this is true, is the whole idea of *learning from experience* — the core of empirical sciences — wrong? Our answer is 'No'. However, we have to be more explicit in what we mean by terms like 'X affects Y', 'X has an effect on Y', 'X influences Y', 'X leads to Y', etc. used in our theories and hypotheses. How can these terms be translated into a language compatible with statistical analyses of empirical data?

Box 1.1 Glossary of New Concepts

$PFE_{xx'}$ — *Prima facie effect* of treatment x compared to treatment x'. It is defined by

$$PFE_{xx'} := E(Y|X{=}x) - E(Y|X{=}x').$$

$PFE_{xx';Z=z}$ — $(Z{=}z)$-*Conditional prima facie effect* of treatment x compared to treatment x'. It is defined by

$$PFE_{xx';Z=z} := E(Y|X{=}x, Z{=}z) - E(Y|X{=}x', Z{=}z).$$

$E(PFE_{xx';Z})$ — *Expectation of the* $(Z{=}z)$-*conditional prima facie effects* of treatment x compared to treatment x'. It is defined by

$$E(PFE_{xx';Z}) := \sum_z PFE_{xx';Z=z} \cdot P(Z{=}z).$$

How to design a study and how to look at the resulting data if we want to probe our theories empirically and learn about the causal dependencies postulated in these theories and hypotheses?

We know that a reversal of total effects does not occur in the randomized experiment, i. e., in an experiment in which observational units (in the social and behavioral sciences, usually the subjects or individuals) are randomly assigned to one of at least two treatment conditions. In the randomized experiment comparing expectation values *is* informative about total causal treatment effects. But why? What is so special in the randomized experiment? Which are the conditions allowing for causal inference in the randomized experiment? Can we create these conditions also in quasi-experimental studies? How can we estimate causal effects in quasi-experiments? And why does randomization not help if we analyze direct treatment effects? Obviously, conclusive answers to these questions can be hoped for only within a theory of causal effects.

Relevance of the Problem

These questions are of fundamental importance for the methodology of empirical sciences and for the empirical sciences themselves. The answers to these questions have consequences for the design and analysis of experiments, quasi-experiments, and other studies aiming at estimating the effects of *treatments*, *interventions*, or *expositions* on certain outcome variables. No *prevention study* can be meaningfully conducted without knowing the concepts of causal effects and how they can be estimated from empirical data, and the same is true for the *evaluation of institutions* such as schools, universities, or clinics with respect to their effects on the outcomes of their clients. Similarly, without a clear concept

of causal effects we are not able to learn from our data about the effects of a certain (possibly harmful) environment on our health, or about the effects of certain behaviors such as smoking or drug abuse. Again, this is similar to the problem of measuring the invisible man's size via the length of his shadow: only with a clear concept of *size*, some basic knowledge in geometry, and the additional information such as the angle of the sun at the time of measurement, are we able to determine his size from the length of his shadow.

Furthermore, without an explicit theory of causal effects we are not able to study direct and indirect effects, and *this is true even in a perfect randomized experiment*. For example, if we are interested in whether or not the effect of vaccination is completely transmitted through the amount of a certain type of antibodies, then this cannot be done relying only on the benefits of a perfect randomized trial. Instead we have to apply certain adjustment techniques. In terms of our metaphor, the 45° angle (the randomized experiment) does not help in determining the parameters we are looking for (the direct effects).

Research Traditions

Of course, raising these questions and attempting answers is not new. Immense knowledge and wisdom about experiments and quasi-experiments has been collected in the Campbellian tradition of experiments and quasi-experiments (see, e. g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish et al., 2002). In the last decades, a more formal approach has been developed supplementing the Campbellian theory and terminology in important aspects: the theory of causal effects in the Neyman-Rubin tradition (see, e. g., Splawa-Neyman, 1923/1990; Rubin, 1974, 2005). Many papers and books indicate the growing influence of this theory (see, e. g., Greenland, 2000, 2004; Höfler, 2005; Rosenbaum, 2002; Rubin, 2006; Winship & Morgan, 1999; Morgan & Winship, 2007) and formidable efforts have already been made to integrate it into the Campbellian framework (West, Biesanz, & Pitts, 2000). Furthermore, these questions have also been dealt with in the graphical modeling tradition (see, e. g., Pearl, 2009; Spirtes et al., 2000) as well as in biometrics, econometrics, psychometrics, and other fields dealing with the methodology of empirical research fields.

Outlook

In this book, we present the theory of total, direct, and indirect causal effects in terms of classical probability theory. We show that a number of questions that have been debated controversially and inconclusively can now be given a clear-cut answer. What kinds of causal effects can be meaningfully defined? Which design techniques guarantee unbiased estimation of causal effects? How to analyze nonorthogonal ANOVA designs (cf., e. g., Aitkin, 1978; Appelbaum & Cramer, 1974; Gosslee & Lucas, 1965; Maxwell & Delaney, 2004; Overall et al., 1975)? How to analyze non-equivalent control-group designs (cf., e. g., Reichardt, 1979)? Should we compare pre-post differences between treatment groups (cf., e. g.,

Lord, 1967; Senn, 2006; van Breukelen, 2006; Wainer, 1991)? Should we use analysis of covariance to adjust for differences in treatment and control that already existed prior to treatment (cf., e. g., Maxwell & Delaney, 2004; Cohen, Cohen, West, & Aiken, 2003)? Should we use new techniques such as propensity score methods instead of the more traditional procedures mentioned above (cf., e. g., Rosenbaum & Rubin, 1984)? How do we deal with non-compliance to treatment assignment (cf., e. g., Cheng & Small, 2006; Dunn et al., 2003; Jo, 2002a, 2002b, 2002c; Jo, Asparouhov, Muthén, Ialongo, & Brown, 2008; J. Robins & Rotnitzky, 2004; J. M. Robins, 1998)? How to analyze direct and indirect effects? We do not treat the statistical sampling models with their distributional assumptions, their implications for parameter estimation, and the evaluation (or tests) of hypotheses about these parameters. However, in chapter 13 we discuss the virtues and problems of general strategies of data analysis such as the analysis of difference scores, analysis of covariance, its generalizations, analysis based on propensity scores, and instrumental variables.

1.5 Exercises

▷ **Exercise 1-1** Why do we need the concept of a causal treatment effect?

▷ **Exercise 1-2** What is the relationship between the unconditional prima facie effect PFE_{10} and the expectations $E(Y|X=0)$ and $E(Y|X=1)$ of the outcome variable Y in the two treatment conditions?

▷ **Exercise 1-3** Verify that Table 1.1 (p. 2) is in fact obtained by collapsing the two corresponding tables for males and females (see Table 1.2, p. 4).

▷ **Exercise 1-4** Which are the three kinds of prima facie effects treated in this chapter?

▷ **Exercise 1-5** What is the difference between statistical inference and causal inference?

▷ **Exercise 1-6** Why are the conditional expectation values $E(Y|X=x)$ in treatment conditions x also probabilities for $Y=1$ in the first example treated in this chapter?

▷ **Exercise 1-7** Compute the conditional probability $P(Y=1 \mid X=1, Z=0)$ from Table 1.2 (p. 4).

▷ **Exercise 1-8** Compute the probability $P(Y=1|X=0)$ of success in the control condition.

▷ **Exercise 1-9** What are the unconditional prima facie effects of the treatments, i. e., the prima facie effects $E(Y|X=1)-E(Y|X=0)$ and $E(Y|X=2)-E(Y|X=0)$ in the second example of this chapter?

▷ **Exercise 1-10** What are the conditional prima facie effects of the treatments, i. e., the prima facie effects $E(Y|X=1,Z=z)-E(Y|X=0,Z=z)$ and $E(Y|X=2,Z=z)-E(Y|X=0,Z=z)$ in the second example of this chapter?

▷ **Exercise 1-11** What are the averages of the conditional prima facie effects

$$E(Y|X=1,Z=z) - E(Y|X=0,Z=z) \qquad \text{and} \qquad E(Y|X=2,Z=z) - E(Y|X=0,Z=z)$$

in the second example of this chapter?

▷ **Exercise 1-12** Compute the coefficients of the equation for the conditional expectation $E(Y|X,M)$ presented in Equation (1.3).

▷ **Exercise 1-13** Compute the coefficients of the equation for the conditional expectation $E(Y|X,M,Z,W)$ presented in Equation (1.4).

▷ **Exercise 1-14** Download *table.1.4.10000.sav* from *www.causal-effects.de*. This data set has been generated from Table 1.4 (p. 8) for a sample of size $N = 10.000$.

(a) Estimate the cell means and the relative frequencies of observations in each of the nine cells of the 3×3 table.

(b) Use each of the procedures offered by your statistical program package to analyze the data including a test of the main effects of the treatment factor (most programs offer Typ I, II and III sums of squares for such an analysis).

(c) Compare the results of these analyses to the parameters presented in Table 1.4 (p. 8).

▷ **Exercise 1-15** Download *table.1.6.10000.sav* from *www.causal-effects.de*. This data set has been generated from Table 1.6 (p. 12) for a sample of size $N = 10.000$.

(a) Estimate the conditional expectation of Y given X and M.

(b) Estimate the conditional expectation of Y given X, M, Z and W.

(c) Compare the estimated regression coefficients to the parameters presented in Equations (1.3) and (1.4), respectively.

Solutions

▷ **Solution 1-1** We need the concept of a causal treatment effect, because Simpson's paradox shows that differences between expectations are meaningless for the evaluation of the effects of a treatment, unless we can show how the differences between expectations are related to the causal effects. Without a definition of causal treatment effects, this would not be possible. Estimating causal treatment effects is crucial for answering questions such as 'Does the treatment help our patients with respect to the outcome variable considered?'

▷ **Solution 1-2** The unconditional prima facie effect PFE_{10} is defined as the difference between the two expectations $E(Y|X=1)$ and $E(Y|X=0)$.

▷ **Solution 1-3** This can easily be verified by adding the probabilities for the observations of the pairs (x,z) of X and Z over males and females. This yields $.144 + .096 = .240$, $.004 + .228 = .232$, $.336 + .024 = .360$ and $.016 + .152 = .168$.

▷ **Solution 1-4** The three kinds of prima facie effects treated in this chapter are: the *unconditional prima facie effect*, the *conditional prima facie effect* given the value z of a covariate Z, and the *average of the* $(Z=z)$*-conditional prima facie effects*. The unconditional prima facie effect of treatment 1 compared to treatment 0 is the difference $PFE_{10} := E(Y|X=1) - E(Y|X=0)$ between the expectations of an outcome variable Y in the two treatment conditions. The $(Z=z)$-conditional prima facie effect is the difference $PFE_{10;Z=z} := E(Y|X=1,Z=z) - E(Y|X=0,Z=z)$ between the $(Z=z)$-conditional expectations of the outcome variable Y in the two treatment conditions. The average prima facie effect is the expectation of the conditional prima facie effects [see Eq. (1.2)].

▷ **Solution 1-5** In *statistical* inference we estimate and test hypotheses about parameters characterizing the distribution of a random variable from sample data. In *causal* inference we interpret some of these parameters as causal effects.

▷ **Solution 1-6** $E(Y|X{=}x) = P(Y{=}1|X{=}x)$, because, in this example, Y is dichotomous with values 0 and 1. In this case, $E(Y|X{=}x) := \sum_y y \cdot P(Y{=}y|X{=}x)$ [see Steyer & Nagel, in press-a, Eq. (9.19)] yields $E(Y|X{=}x) = 0 \cdot P(Y{=}0|X{=}x) + 1 \cdot P(Y{=}1|X{=}x) = P(Y{=}1|X{=}x)$.

▷ **Solution 1-7** According to Table 1.2 (p. 4) ,

$$P(Y{=}1|X{=}1,Z{=}0) = \frac{P(X{=}1,Y{=}1,Z{=}0)}{P(X{=}1,Z{=}0)} = \frac{.016}{.016+.004} = .80.$$

▷ **Solution 1-8** First of all, note that the theorem of total probability, can also be applied to conditional probabilities, in this exercise, the $(X{=}0)$-conditional probabilities. Hence, according to this theorem,

$$P(Y{=}1|X{=}0) = P(Y{=}1|X{=}0,Z{=}0)\cdot P(Z{=}0|X{=}0) + P(Y{=}1|X{=}0,Z{=}1)\cdot P(Z{=}1|X{=}0).$$

The probabilities $P(Y{=}1|X{=}0,Z{=}0) = .70$ and $P(Y{=}1|X{=}0,Z{=}1) = .20$ are computed analogously to Exercise 1-7 and the other two probabilities occurring in this formula are $P(Z{=}0|X{=}0) = .48/.60$ and $P(Z{=}1|X{=}0) = .12/.60$ (see Table 1.2, p. 4). Hence,

$$P(Y{=}1|X{=}0) = \frac{.70\cdot.48}{.60} + \frac{.20\cdot.12}{.60} = .60.$$

▷ **Solution 1-9** The prima facie effects $E(Y|X{=}1) - E(Y|X{=}0)$ and $E(Y|X{=}2) - E(Y|X{=}0)$ can be computed from Table 1.3 (p. 7). They are as follows:

$$PFE_{10} = E(Y|X{=}1) - E(Y|X{=}0) = 100.00 - 111.25 = -11.25$$

and

$$PFE_{20} = E(Y|X{=}2) - E(Y|X{=}0) = 114.25 - 111.25 = 3.00.$$

▷ **Solution 1-10** The conditional prima facie effects $E(Y|X{=}1,Z{=}z) - E(Y|X{=}0,Z{=}z)$ and $E(Y|X{=}2,Z{=}z) - E(Y|X{=}0,Z{=}z)$ can be computed from Table 1.4 (p. 8). For *low neediness* $(Z{=}0)$, they are:

$$PFE_{10;Z=0} = E(Y|X{=}1,Z{=}0) - E(Y|X{=}0,Z{=}0) = 100 - 120 = -20$$

$$PFE_{20;Z=0} = E(Y|X{=}2,Z{=}0) - E(Y|X{=}0,Z{=}0) = 80 - 120 = -40.$$

For *medium neediness* $(Z{=}1)$, they are:

$$PFE_{10;Z=1} = E(Y|X{=}1,Z{=}1) - E(Y|X{=}0,Z{=}1) = 100 - 110 = -10$$

$$PFE_{20;Z=1} = E(Y|X{=}2,Z{=}1) - E(Y|X{=}0,Z{=}1) = 90 - 110 = -20.$$

Finally, for *high neediness* $(Z{=}2)$, the conditional prima facie effects are:

$$PFE_{10;Z=2} = E(Y|X{=}1,Z{=}2) - E(Y|X{=}0,Z{=}2) = 100 - 60 = 40$$

$$PFE_{20;Z=2} = E(Y|X{=}2,Z{=}2) - E(Y|X{=}0,Z{=}2) = 140 - 60 = 80.$$

▷ **Solution 1-11** Using the results of the last exercise, the average of the $(Z{=}z)$-conditional prima facie effects can be computed from the conditional effects as follows:

$$\begin{aligned} E(PFE_{10;Z}) &= PFE_{10;Z=0}\cdot P(Z{=}0) + PFE_{10;Z=1}\cdot P(Z{=}1) + PFE_{10;Z=2}\cdot P(Z{=}2) \\ &= -20\cdot\frac{1}{4} - 10\cdot\frac{1}{2} + 40\cdot\frac{1}{4} = 0. \end{aligned}$$

$$\begin{aligned} E(PFE_{20;Z}) &= PFE_{20;Z=0}\cdot P(Z{=}0) + PFE_{20;Z=1}\cdot P(Z{=}1) + PFE_{20;Z=2}\cdot P(Z{=}2) \\ &= -40\cdot\frac{1}{4} - 20\cdot\frac{1}{2} + 80\cdot\frac{1}{4} = 0. \end{aligned}$$

▷ **Solution 1-12** The two coefficients $\beta_1 \approx -3.7528$ and $\beta_2 \approx 1.1876$ are obtained by

$$\begin{aligned} \boldsymbol{\beta} = \boldsymbol{\Sigma}_{VV}^{-1}\boldsymbol{\Sigma}_{Vy} &= \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} \approx \begin{pmatrix}0.25 & 5.00\\ 5.00 & 189\end{pmatrix}^{-1}\begin{pmatrix}5.00\\ 205.70\end{pmatrix} \\ &\approx \begin{pmatrix}8.4944 & -0.2247\\ -0.2247 & 0.0112\end{pmatrix}\begin{pmatrix}5.00\\ 205.70\end{pmatrix} \approx \begin{pmatrix}-3.7528\\ 1.1876\end{pmatrix} \end{aligned}$$

[see Steyer & Nagel, in press-a, Eq. (12.54)]. The appropriate statements in R are:

```
a=matrix(c(.25,5,5,189),byrow=T,nrow=2,ncol=2)
b=matrix(c(5,205.7),byrow=T,nrow=2,ncol=1)
round(solve(a,b),4)
```

In this equation, $\boldsymbol{\Sigma}_{VV}^{-1}$ denotes the inverse of the covariance matrix of $V := (X, M)$ and $\boldsymbol{\Sigma}_{Vy}$ the covariance vector of $V = (X, M)$ and Y. The intercept $\beta_0 \approx 34.989$ is obtained by

$$\begin{aligned} \beta_0 &\approx E(Y) - \beta_1\cdot E(X) + \beta_2\cdot E(M) \\ &\approx E(Y) + 3.7528\cdot E(X) - 1.1876\cdot E(M) \\ &\approx 140 + 3.7528\cdot 0.50 - 1.1876\cdot 90 \approx 34.9924 \end{aligned}$$

[see Steyer & Nagel, in press-a, Eq. (12.53)].

▷ **Solution 1-13** The coefficients γ_1 to γ_4 of

$$E(Y|X, M, Z, W) = \gamma_0 + \gamma_1\cdot X + \gamma_2\cdot M + \gamma_3\cdot Z + \gamma_4\cdot W$$

are obtained by

$$\begin{aligned} \boldsymbol{\gamma} = \boldsymbol{\Sigma}_{RR}^{-1}\boldsymbol{\Sigma}_{Ry} &= \begin{pmatrix}\gamma_1\\ \gamma_2\\ \gamma_3\\ \gamma_4\end{pmatrix} = \begin{pmatrix}0.25 & 5.00 & 0.00 & 0.00\\ 5.00 & 189 & 80 & 68\\ 0.00 & 80 & 100 & 85\\ 0.00 & 68 & 85 & 100\end{pmatrix}^{-1}\begin{pmatrix}5.00\\ 205.70\\ 116.50\\ 124.00\end{pmatrix} \\ &= \begin{pmatrix}20.0000 & -0.8000 & 0.6400 & 0.0000\\ -0.8000 & 0.0400 & -0.0320 & 0.0000\\ 0.6400 & -0.0320 & 0.0616 & -0.0306\\ 0.0000 & 0.0000 & -0.0306 & 0.0360\end{pmatrix}\begin{pmatrix}5.00\\ 205.70\\ 116.50\\ 124.00\end{pmatrix} = \begin{pmatrix}10.00\\ 0.50\\ 0.00\\ 0.90\end{pmatrix} \end{aligned}$$

[see again Steyer & Nagel, in press-a, Eq. (12.54)]. In this equation, $\boldsymbol{\Sigma}_{RR}^{-1}$ denotes the inverse of the covariance matrix of $R := (X, M, Z, W)$ and $\boldsymbol{\Sigma}_{Ry}$ the covariance vector of $R = (X, M, Z, W)$ and Y. The appropriate statements in R are:

```
a=matrix(c(.25,5,0,0,5,189,80,68,0,80,100,85,0,68,85,100),
         byrow=T,nrow=4,ncol=4)
b=matrix(c(5,205.7,116.5,124),byrow=T,nrow=4,ncol=1)
round(solve(a,b),4).
```

The intercept $\gamma_0 = 0.00$ is obtained by

$$\begin{aligned}\gamma_0 &= E(Y) - [\gamma_1 \cdot E(X) + \gamma_2 \cdot E(M) + \gamma_3 \cdot E(Z) + \gamma_4 \cdot E(W)] \\ &= E(Y) - 10 \cdot E(X) - 0.50 \cdot E(M) - 0.00 \cdot E(Z) - .90 \cdot E(W) \\ &= 140 - 10 \cdot 0.50 - 0.50 \cdot 90 - 0.00 \cdot 100 - .90 \cdot 100 \;=\; 0.00.\end{aligned}$$

[see again Steyer & Nagel, in press-a, Eq. (12.53)].

▷ **Solution 1-14** No solution provided. Just compare your results to the parameters presented in Table 1.4 (p. 8).

▷ **Solution 1-15** No solution provided. Just compare your estimated parameters to the true parameters presented in Equations (1.3) and (1.4).

Chapter 2
Some Typical Random Experiments

In chapter 1 we have shown that comparing conditional expectation values of an outcome variable between treatment groups can be completely misleading if used for the evaluation of treatment effects. We have also shown that regression coefficients and the conditional expectations they describe can be completely misleading even in the randomized experiment, if used to determine the direct treatment effect with respect to an intermediate variable M. In this chapter we will prepare the stage for the theory of causal effects, describing the kind of empirical phenomena it refers to: single-unit trials of experiments or quasi-experiments, but also single-unit trials of observational studies in which causal effects can be investigated.

A single-unit trial is a specific random experiment. Note the distinction between a *random experiment* and a *randomized experiment*. Stochastic dependencies between events and between random variables always refer to a random experiment, but not necessarily to a *randomized experiment* in which a subject is assigned to one of the treatment conditions by a randomization procedure. In the simplest case of such a randomization we assign the subject to treatment or control according to the outcome of flipping a coin. In contrast, a *random experiment* is the concrete empirical phenomenon to which stochastic dependencies between events and random variables described by conditional distributions, probabilities, correlations, and conditional expectations refer to.

The single-unit trial *is not the sample* dealt with in statistical models. In a sample, the single-unit trial is repeated many times in one way or another. This is necessary when it comes to estimating parameters and testing hypotheses about these parameters, some of which might be causal effects. The single-unit trial does *not allow* treating problems of parameter estimation or hypothesis testing. However, it is sufficient for defining causal effects and studying how to identify them, i. e., studying under which conditions and how they can be computed from empirically estimable parameters.

A single-unit trial is also what we refer to in substantive hypotheses and theories. Furthermore, single-unit trials are what is of interest in practical work. How does the treatment of a patient affect the outcome of this patient if compared to another possible treatment? What is the treatment effect for a male, and what is its effect for a female? What is the direct treatment effect (e. g., of vaccination) on the outcome variable (e. g., influenza) that is *not* transmitted through a specific intermediate variable (e. g., a measure of certain antibodies)? Which variables ex-

plain inter-individual differences in individual causal effects? All these questions are raised using concepts referring to single-unit trials.

Overview

We start with the single-unit trial of simple experiments and then treat increasingly more complex ones introducing additional design features. Specifically, we will introduce the single-unit trials of experiments and quasi-experiments with fallible covariates, a multifactorial design with more than one treatment, multilevel experiments and quasi-experiments, and experiments and quasi-experiments with intermediate variables and latent outcome variables.

We also discuss different kinds of random variables that will play a crucial role in the chapters to come. Among these random variables are the observational-unit variable, other manifest and latent covariates, treatment variables, intermediate variables, as well as manifest and latent outcome variables. In this chapter, we confine ourselves to an informal description of single-unit trials and the random variables involved, preparing the stage for their mathematical representations in the following chapters.

2.1 Simple Experiments

As a first class of random experiments, we consider the single-unit trials of *simple experiments and quasi-experiments*. Such single-unit trials are between-subjects experiments and quasi-experiments in which *no fallible covariates* are assessed.

Such a single-unit trial consists of:

(a) sampling an observational unit u (e. g., a person) from a set (sometimes called 'population') of units,
(b) assigning the unit or observing its assignment to one of several experimental conditions (represented by the value x of the treatment variable X),
(c) recording the numerical value y of the outcome variable Y.

Figure 2.1 displays a tree representation of the set of possible outcomes of this single-unit trial. Note that this is the kind of random experiment we (implicitly) referred to describing Simpson's paradox in chapter 1. The random variables X (treatment), Y (success), and Z (sex), the conditional expectation values $E(Y|X{=}x)$ and $E(Y|X{=}x,Z{=}z)$, as well as the probabilities $P(X{=}x)$, $P(Z{=}z)$, $P(X{=}x,Z{=}z)$ all referred to such a single-unit trial. Of course, all these conditional expectation values and probabilities are unknown in applications. Nevertheless, they are the parameters that stochastically determine the outcome of the single-unit trial, just in the same way as the probability of tossing *heads* stochastically determines the outcome of flipping a coin.

In order to illustrate this point, imagine flipping a deformed coin that has the shape of a Chinese wok and suppose that in this case the probability of flipping *heads* is .80 instead of .50. Although these probabilities do not deterministically

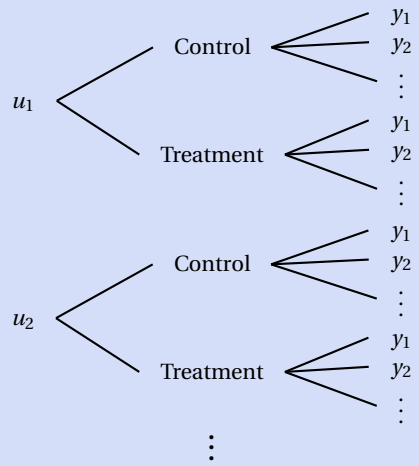

Figure 2.1. A simple experiment or quasi-experiment

determine the outcomes of flipping the coins, they stochastically determine the outcomes.

In fact, we may consider the single-unit trial of (a) sampling a coin u from a set of coins, (b) forming ($X=1$) or not forming ($X=0$) a wok out of it, and (c) observing whether ($Y=1$) or not ($Y=0$) we then toss *heads*. In this single-unit trial, the difference .80 − .50 = .30 would be the causal effect of the treatment variable X on the outcome variable Y. Let us emphasize that the probabilities .80 and .50 and their difference .30 refer to this single-unit trial, although these probabilities can only be estimated if we conduct many of these single-unit trials, i. e., if we draw a sample. However, if these probabilities were known, we could dispense with a sample and the data that would result from drawing it (see Exercise 2-1), and still have a perfect prediction for the outcome of such a single-unit trial.

2.1.1 Sampling a Unit

The first part of this single-unit trial consists of sampling an observational unit. In the social sciences, units often are persons, but they might be groups, school classes, schools and even countries. Usually such units change over time. Therefore, it should be emphasized that, in simple experiments and quasi-experiments, we are talking about the *units at the onset of treatment*. Later we will see that we have to distinguish between *units at the onset of treatment* and *units at the time of assessment of the outcome variable*, which might be months or even years later. In a single-unit trial of simple experiments and quasi-experiments, the units can

be represented by the observational-unit variable U, whose possible values u are the *units at the onset of treatment*.

Note that the unit at the onset of treatment also comprises his or her experiences a year and/or the day before treatment, as well as the psycho-bio-social situation in which he or she is at the onset of treatment. Both, the experiences and the situation, already happen *before* the onset of treatment. Therefore, they are attributes of the observational units u, and this is true although they once were just possible events that had some (unknown) probabilities to occur. Looking at them at the time of the onset of treatment, they are no events any more that may or may not occur. Instead, these prior experiences and situations are then *fixed attributes* of the units. They can be treated in the same way as other attributes such as sex and educational status.

2.1.2 Treatment Variable

In an experiment or quasi-experiment, there is always a treatment variable, which we usually denote by X. The unit drawn is either assigned — e. g., by the experimenter or by some other person (such as a physician, a psychologist, or a social worker) — to one of the possible treatments, or we observe self-selection to one of the treatment conditions. In the simplest case there are at least two treatment conditions, e. g., *treatment* and *control*. These treatment conditions are the possible values of the treatment variable X. For simplicity, we use the values $0, 1, \ldots, J$ to represent $J+1$ treatment conditions. Furthermore, unless stated otherwise, we presume that treatment *assignment* and actual *exposure* to treatment will be equivalent, i. e., unless stated otherwise, we assume that there is perfect compliance.

Selection of a unit into one of the treatment conditions x may happen with unknown probabilities, e. g., when there is self-selection or assignment by an unknown physician. This is often the case in quasi-experiments. However, assignment can also be done with known probabilities that are equal for different units (such as in the simple randomized experiment) or with known probabilities that may be unequal for different units (such as in the conditionally randomized experiment). In this case, these treatment probabilities may also depend on a covariate Z representing pre-treatment attributes of the units. *Conditional and unconditional randomized assignment, distinguish the true experiment from the quasi-experiment*, in which the assignment or selection probabilities are unknown. (See section 7.5 for more details on randomization and conditional randomization.)

2.1.3 Covariates

In simple experiments and quasi-experiments, the focus is usually on the treatment effects on an outcome variable. Hence, if we are interested in the treatment variable as a cause, then each attribute of the observational units is a covariate. Examples are *sex*, *race*, *educational status*, and *socio-economic status*. Once

the unit is drawn, its *sex, race, educational status,* and *socio-economic status* are fixed as well. This means that there is no additional sampling process associated with assessing these covariates. This is also the reason why they do not appear in points (a) to (c) describing the single-unit trial (see p. 24).

Because covariates represent attributes of the unit *at the onset of treatment* they can never be affected by the treatment. However, there can be (stochastic) dependencies between the treatment variable and covariates. In Simpson's paradox, for instance, there is a strong correlation between *sex* and the treatment variable.

Multidimensional Covariates

Covariates may be uni- or multi-dimensional, qualitative (such as $Z_1 :=$*sex* and $Z_2 :=$ *educational background*) or quantitative (such as $Z_3 :=$ *height* and $Z_4 :=$*body mass index*) or, if it is a multivariate variable made up of several uni-dimensional variables, it may consist of qualitative *and* quantitative covariates such as $Z_5 := (Z_1, Z_4)$.

Specific Covariates

Note that the U-conditional treatment probabilities $P(X{=}x|U)$ and the Z-conditional treatment probabilities $P(X{=}x \mid Z)$ are covariates as well, provided that Z is a covariate. (The mathematical background for this statement are chapters 2 and 10 of Steyer and Nagel (in press-a).) Furthermore, the *assignment* to treatment x with values 'yes' and 'no' is also covariate, if *assignment to treatment* and *exposure to treatment* (again with values 'yes' and 'no') are not identical. This distinction is useful in experiments with non-compliance (see, e. g., Jo, 2002a, 2002b, 2002c; Jo et al., 2008).

Unobserved Covariates

Even if we consider a multivariate covariate Z consisting of several univariate covariates, there are always unobserved variables that are prior or simultaneous to treatment. Such variables are called *unobserved covariates*. Sometimes they are also called *hidden confounders* (cf., e. g., Rosenbaum, 2002). Of course such an unobserved covariate may bias the conditional expectation values of the outcome variable just in the same way as an observed covariate. Whether or not the conditional expectation values of the outcome variable in the treatment conditions are unbiased such that their differences represent causal effects does not only depend on the relationship between the measured variables such as X, Y, and the observed (possible multivariate) covariate, say Z, but also on the relationship of these variables to the unobserved covariates. In other words, covariates exert their maleficent effects irrespective of whether or not we observe them.

2.1.4 Outcome Variable

Of course, the outcome variable Y refers to a time at which the treatment might have had its impact. Hence, treatment variables are always prior to the outcome variable. In principle, we may also observe several outcome variables, e. g., in order to study how effects of a treatment grow or decline over time or to study effects that are not limited to a single outcome variable. All random variables mentioned above refer to a concrete single-unit trial and they have a joint distribution. Each combination of unit, treatment condition, and score of the outcome variable may be an observed result of such a single-unit trial. This implies that the variables U, Z, X, and Y, as well as unobserved covariates, say W, have a joint distribution. (See, e. g., section 5.3 of Steyer and Nagel (in press-a).) Once we specified the random experiment to be studied, this joint distribution is fixed, even though it might be known only in parts or even be unknown completely.

2.1.5 Causal Effects and Causal Dependencies

There is already a plenitude of different kinds of causal effects and causal dependencies that can be considered in the single-unit trial of a simple experiment or quasi-experiment. For simplicity, suppose the treatment has just two values, say *treatment* and *control*. *First*, there is the *average total effect* of treatment (compared to control) on the outcome variable Y. *Second*, there are the *conditional total treatment effects* on Y, where we may condition on any function of the observational-unit variable U. If, e. g., $Z :=$ *sex* with values m for *male* and f for *female*, then we may consider the $(Z=m)$-conditional total treatment effect on Y, i. e., the average total treatment effect for males, and the $(Z=f)$-conditional total treatment effect on Y, i. e., the average total treatment effect for females. Similarly, if $Z :=$ *socio-economical status*, we may consider the conditional total treatment effects on Y for each status group, etc. *Third*, although difficult and often impossible to estimate, we may also consider the *individual total effect* of *treatment* compared to *control* on Y.

By definition, within a *simple* experiment and quasi-experiment we cannot consider any *direct* treatment effects with respect to a specified intermediate variable, i. e., the effects of the treatment on the outcome variable that *are not* transmitted through a specified intermediate variable M. However, the total treatment effects discussed above are, of course, transmitted through intermediate variables, irrespective of whether or not we observe (or are aware of) these intermediate variables. (See section 2.5 for experiments and quasi-experiments with observed intermediate variables).

Aside from the treatment effects discussed above we can also consider the causal effects of a covariate. Among the causal effects of such a covariate are its average total effect on the outcome variable Y, its conditional direct effects on Y given the different values of the treatment variable — which now takes the role of an intermediate variable — the average of these X-conditional direct effects and its indirect effect mediated by X. In a randomized experiment, e. g., the

causal effects of all covariates on X will be zero. In other words, the zero prima facie effects created by randomization will *not be biased* or *spurious*. In contrast, in quasi-experiments, causal effects of some covariates on X might be different from zero. In self-selection, e. g., *neediness* for a therapy might have strong average effects on the treatment variable. Furthermore, neediness often has strong $(X{=}x)$-conditional direct effects and a strong average direct effect on the outcome variable if it measures some aspects of health.

These effects of the covariates on the treatment variable and on the outcome variable are discussed in the literature on structural equation modeling (see, e. g., Bentler, 1995; Bollen, 2002; Kaplan, 2000; S.-Y. Lee, 2007; Little, Card, Bovaird, Preacher, & Crandall, 2007; MacCallum & Austin, 2000; Muthén & Muthén, 1998-2007) and graphical modeling (see, e. g., Cox & Wermuth, 2004; Greenland, Pearl, & Robins, 1999; Spirtes et al., 2000; Pearl, 1995, 1998, 2009), whereas they have been criticized in the Rubin tradition (see, e. g., Holland, 1986). What should be noted is that the causal effects of covariates have no 'individual causal interpretation' (see, e. g., Borsboom, Mellenbergh, & van Heerden, 2003). While the average total effect of a treatment variable can be interpreted as the effect of the treatment on an unknown, randomly drawn unit that can be exposed to treatment or control, the effects of a covariate such as *neediness* or *sex* on an outcome variable do *not* have such an individual causal interpretation. The unit (at the onset of treatment) *has* a certain degree of *neediness* and it *has* a sex, but it cannot be *exposed* to a neediness condition or a sex. Nevertheless, neediness effects and sex effects can be 'spurious' or 'biased', and we can define and aim at estimating 'unbiased' or 'causal' neediness and sex effects. As we discussed in more detail in chapter 5, we can consider both, a treatment variable or an attribute of the units, as causes. The conceptual framework provided in the chapters to come will cover both kinds of causal effects which seem — and with respect to manipulability at the individual level *are* — different from a content point of view.

2.2 Experiments With Fallible Covariates

Another class of random experiments are single-unit trials of experiments and quasi-experiments in which we assess a fallible covariate. In this case, there is at least one covariate that is *not* a (deterministic) attribute of the observational units. The single-unit trial of such an experiment of quasi-experiment consists of:

(a) sampling an observational unit u (e. g., a person) from a population of units,
(b) assessing the values $z_1, \ldots, z_k$ of the covariates $Z_1, \ldots, Z_k$, $k \geq 1$.
(c) assigning the unit or observing its selection to one of several experimental conditions (represented by the value x of the treatment variable X),
(d) recording the numerical value y of the outcome variable Y.

The crucial difference to a simple (quasi-) experiment is that there is variability of the manifest covariate *given the observational unit u* (see Fig. 2.2). In this case,

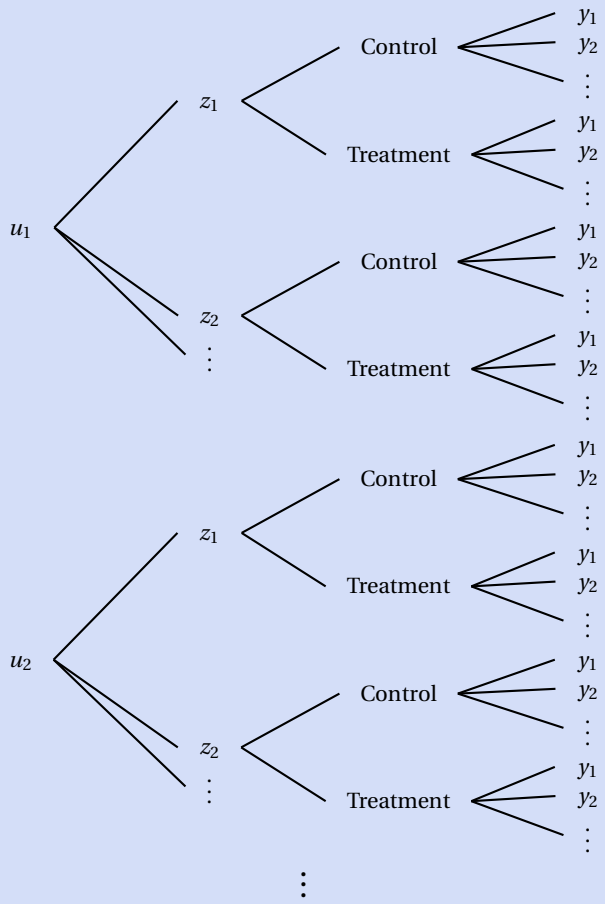

Figure 2.2. Experiment or quasi-experiment with a fallible covariate

we may distinguish between the *latent covariate*, say ξ, representing the attribute to be assessed and its *fallible measures*, the manifest variables $Z_1, \ldots, Z_k$ actually observed. This distinction does not only open up the possibilities to study the effects of the latent covariate on the treatment variable and on the outcome variable, but also for investigating whether or not the dependency of the fallible measures on the latent variable is causal.

Furthermore, this distinction also implies that the *unit* whose attributes are measured *at the time when the covariate is assessed* is not identical any more with the *unit at the onset of treatment* (see section 2.1). The covariate might be assessed some months before the treatment is given — enough time and plenty of possibilities for the unit to change in various ways, e. g., due to maturation,

learning, critical life events, and other experiences that are not fixed at the time of assessing the covariate. As a consequence, a variable, say W, representing such intermediate events and experiences may also affect the outcome variable Y over and above (a) the covariate Z, (b) the treatment variable X, and (c) the observational-unit variable U, which now represents the *observational units at the time of the assessment of the covariate* $Z := (Z_1, \ldots, Z_k)$. In other words, the outcome variable Y does not necessarily only depend on the units, covariates , and the treatment variable alone. Instead, it may also depend on an unobserved covariate W lurking in between the assessment of the observed covariate Z and the onset of treatment. This is one of the reasons why we need to define causal effects in a more general way than in the Neyman-Rubin tradition (see chs. 4 and 5).

Note that assessing a fallible covariate does not only change the interpretation of the observational-unit variable U, but it also changes the random experiment, and with it, the empirical phenomenon we are considering. Assessing, prior to treatment, a fallible covariate such as a pre-test of an ability, an attitude, or a personality trait, may change the observational units and their attributes, as well the effects of the treatment on a specified outcome variable, which usually is related to such a pre-test. This has already been discussed by Campbell and Stanley (1963), who also recommended designs for studying the effects of pre-treatment assessment.

Covariates

What are the covariates in such a single-unit trial? First of all, we have to choose the cause to be considered. If it is the treatment variable X, then each attribute of the unit at the time of the assessment of the observed covariates $Z_1, \ldots, Z_k$ is a covariate pertaining to X as well. This does not only include variables such as *sex*, *race*, and *educational status*, but also the latent covariate, say ξ, (which might be multi-dimensional). Furthermore, aside from the manifest covariates $Z_1, \ldots, Z_k$, each variable W representing an intermediate event or experience of the unit (occurring in between the assessment of the observed covariates and the onset of the treatment), as well as any attribute of the *unit at the onset of treatment* is a covariate as well, irrespective of whether or not these covariates are observed.

Note that a latent covariate ξ may be considered a cause of its fallible measures $Z_1, \ldots, Z_k$ but also of the outcome variable Y. This is not in conflict with the theory that the treatment variable X is a cause of Y as well. In this kind of single-unit trial, we have several causes and several outcome variables that are affected by these causes. Again it would be possible to consider the treatment variable X to be causally dependent on the manifest or latent covariates. In other words, we may also raise the question if the treatment probabilities $P(X{=}1 \mid Z_1, \ldots, Z_k)$ or $P(X{=}1 \mid \xi)$ describe causal dependencies. This makes clear that the term 'covariate' can only be defined with respect to a focused cause.

2.3 Two-Factorial Experiments

As a third class of random experiments we consider two-factorial experiments. The single-unit trial of such a two-factorial experiment or quasi-experiment consists of:

(a) sampling an observational unit u (e. g., a person) from a population of units,
(b) assigning the unit or observing its assignment to one of several experimental conditions that are defined by the pair (x, z) of levels of two treatment variables X and Z, respectively.
(c) recording the numerical value y of the outcome variable Y.

Sampling a Unit

Because we presume that no fallible covariates such as 'severity of symptoms', 'motivation for treatment', etc. are assessed before treatment, sampling an observational unit means that we are sampling a *unit at the onset of treatment*.

Treatment Variables

As a simple example, let us consider an experiment in which we study the effects — including the joint effects — of two treatment factors, say *individual therapy* represented by X (with values 'yes' and 'no') and *group therapy* represented by Z (with values 'yes' and 'no').

In such a two-factorial experiment, we consider *group therapy* as a covariate and *individual therapy* to be the treatment variable in order to ask for the conditional and average total effects of individual therapy given group therapy. In contrast, we may also consider *individual therapy* to be a covariate and *group therapy* to be the focused treatment variable. Finally, we may also consider the two-dimensional variable (X, Z) as the treatment variable. Which option is chosen depends on the causal effects we are interested in (see below).

Outcome Variable

Again, the outcome variable Y refers to a time at which the treatment might have had the impact to be estimated. Hence, both treatment variables are prior to the outcome variable considered. And again, we may also observe several outcome variables, e. g., in order to study how effects of a treatment grow or decline over time or to study effects that are not limited to a single outcome variable.

Causal Effects

There are several causal effects we might look at. If X and Z have only two values, then we may be interested in the following effects on the outcome variable Y:

(a_1) the conditional total effect of 'individual therapy' as compared to 'no individual therapy' given that the unit treated also receives 'group therapy',
(b_1) the corresponding conditional total effect given that the unit does *not* receive 'group therapy', and
(c_1) in the average of these conditional total effects of 'individual therapy' as compared to 'no individual therapy', averaging over the two values of Z.

Vice versa, we might also be interested in the following effects on the outcome variable Y:

(a_2) the conditional total effect of 'group therapy' as compared to 'no group therapy' given that the unit treated also receives 'individual therapy',
(b_2) the corresponding conditional total effect given that the unit does *not* receive 'individual therapy', and
(c_2) in the average of these conditional total effects of 'group therapy' as compared to 'no group therapy', averaging over the two values of X.

Furthermore, there are other causal effects on Y we might study, namely

(a_3) the total effect of receiving 'individual therapy' *and* 'group therapy' as compared to receiving none of the two treatments.
(b_3) the total effect of receiving 'individual therapy' *and* 'group therapy' as compared to receiving 'individual therapy' only.
(c_3) the total effect of receiving 'individual therapy' *and* 'group therapy' as compared to receiving 'group therapy' only.
(d_3) the total effect of receiving 'individual therapy' *and* 'no group therapy' as compared to receiving 'group therapy' *and* 'no individual therapy'.

All these effects may answer meaningful causal questions and in fact, there are even more causal effects than those listed above even if we do not count the various conditional total effects we might want to study if additional covariates such as *sex* or *educational status* are considered.

Covariates

If we focus the effect of X (individual therapy), then we consider Z (group therapy) as a covariate , whereas we treat X as a covariate if we study the effects of Z (group therapy). In both cases, each attribute of the unit at the onset of treatment (such as *sex* or *educational status*) could be considered as covariates as well. Assessing these covariates does not appear in points (a) to (c) of the random experiment, because these covariates are (deterministic) functions of the observational-unit variable. Therefore, there is no additional sampling process associated with their assessment.

This is also true for other covariates, e. g., variables characterizing the situation in which the unit is at the onset of treatment, the number of *hours slept* last night, or *day time* at which the unit receives its treatment. Even variables that characterize early experiences in the childhood of the unit such as a *broken home* or

mother's child care behavior are covariates in this single-unit trial. They are there and exert their effects even if they are not assessed.

Note, again that assessment of these covariates in a questionnaire filled in by the person constitutes a new random experiment that may differ in important ways from a random experiment in which the unit has no such task (see section 2.2). In psychology, an assessment often is a treatment of its own.

2.4 Multilevel Experiments

In multilevel experiments and quasi-experiments we also study the effect of a treatment on an outcome variable. However, in such a design, the observational units are nested within higher hierarchical units referred to as *clusters*. Examples include experiments, in which students are nested within classrooms, patients are nested within groups of treated patients, and inhabitants are nested in neighborhoods. Multilevel designs can be classified as designs with treatment assignment at the unit-level or at the cluster-level. Furthermore, multilevel designs differ with respect to the assignment of units to clusters. There are designs with pre-existing clusters and there are designs with assignment of units to clusters. All these designs involve different single-unit trials.

A single-unit trial with *pre-existing clusters* consists of:

(a) sampling a cluster c (e. g., a school class, a neighborhood or a hospital) from a set of clusters,
(b) sampling an observational unit u (e. g., a person) from a set of units within the cluster,
(c) assigning the unit or the cluster (depending on the design) or observing their assignment to one of several experimental conditions (represented by the value x of the treatment variable X),
(d) recording the numerical value y of the outcome variable Y.

In contrast, a *single-unit trial with assignment of units to clusters* consists of:

(a) sampling an observational unit u (e. g., a person) from a population of units,
(b) assigning the unit or observing its assignment to one of several clusters (represented by the value c of the cluster variable C),
(c) assigning the unit or the cluster (depending on the design) or observing their assignment to one of several experimental conditions (represented by the value x of the treatment variable X),
(d) recording the numerical value y of the outcome variable Y.

In the experiment with pre-existing clusters, each unit can only appear in one cluster, whereas in the experiment with assignment of units to clusters, each unit can appear in more than one cluster. Hence, in the latter designs the cluster variable can bias the dependency of the outcome variable on the treatment variable *on the level of the observational unit*. In this aspect this design resembles the multifactorial design described in section 2.3.

Covariates

What are the covariates in multilevel designs if the treatment variable X is considered as the cause? The answer depends on the type of design considered: In designs with treatment assignment at the unit-level, attributes of the observational unit (such as *sex*, *race* or *educational status*) are covariates, but also attributes of the cluster (such as *school type*, *hospital ownership* or cluster-specific expectations of covariates at the unit-level, such as *school-level of socio-economic status* or *school-level intelligence*). In these designs, clusters may not only be considered as covariates, but also as treatments, because some of the effects observed later on may depend on the composition of the group to which a particular unit, say Joe, is assigned. Receiving group therapy together with beautiful Ann in the same group might make a great difference as compared to getting it together with awful Jim. In designs in which clusters as a whole are assigned to treatment conditions, only attributes of the cluster can influence the assignment. Hence, in data analysis we would focus on controlling for the covariates on the cluster level (see, e. g., Nagengast, 2009 for more details).

2.5 Experiments With Intermediate Variables

Another class of random experiments are experiments with *intermediate variables*. The basic goal of such a experiment is to investigate if and to which degree the effect of a cause X (such as a treatment variable) on an outcome variable Y may be *mediated* or *transmitted* by another variable, say M. A first example is mediation of the effect of *vaccination* (with values *yes* or *no*) on *the severity of influenza symptoms* by the *amount of antibodies*. Another example is mediation of the effect of *teachers encouragement* (with values *yes* or *no*) on the *achievement* by the *amount of time spent on learning* (see, e. g., Holland, 1988; Sobel, 2008, or Rubin, 2004).

In the simplest case with a single manifest intermediate variable we consider the following single-unit trial that consists of:

(a) sampling a person u out of a set of persons (the population of persons),
(b) assigning the unit or observing its selection to one of several experimental conditions (represented by the value x of the treatment variable X),
(c) assessing the value m of an intermediate variable M, and
(d) recording the numerical value y of the outcome variable Y.

In this single-unit trial, the values u of the observational-unit variable U again represent the observational *unit at the onset of treatment*, while the intermediate variable M represents some attribute of the *unit at the time point at which the intermediate variable is assessed*. This time point is *in between* the onset of treatment and the assessment of the outcome variable Y (see Fig. 2.3). If M is fallible, then we distinguish between M and the latent variable to be measured by M. In this case we would need an additional layer in the tree representation for the

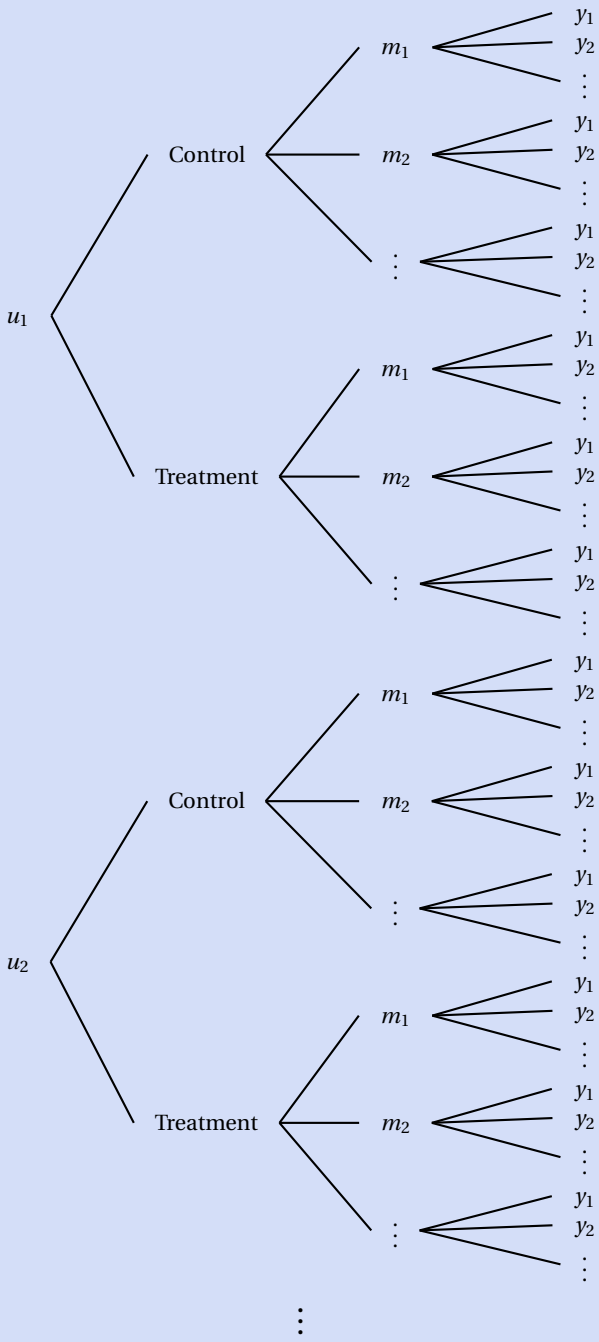

Figure 2.3. Experiment or quasi-experiment with an intermediate variable

latent intermediate variable. Furthermore, instead of a single manifest intermediate variable, we would need several manifest intermediate variables measuring the latent intermediate variable.

Covariates

What are the covariates in such a single-unit trial? Again, the answer depends on the choice of the cause. If it is the treatment variable X, then each attribute of the unit at the onset of treatment is a covariate pertaining to X. Examples are *sex*, *race*, and *educational status*. Note that the set of covariates of X is the same irrespective of the choice of the outcome variable. In this example we may choose the original outcome variable Y; however, we may also choose the intermediate variable M as an outcome variable in order to study the effects of X on M.

Focusing M as a cause, brings additional covariates into play, namely all those variables that are *in between* treatment and the assessment of the intermediate variable. This could be critical life events, additional drugs taken after treatment and before the assessment of the intermediate variable, or an additional treatment to which the unit is exposed and which may or may not be manipulated by the experimenter.

2.6 Experiments With Latent Outcome Variables

We may also consider single-unit trials of experiments with a *latent outcome variable*. The basic goal of such experiments is to investigate the effect of the treatment variable X on a *latent* outcome variable, say η. This is of substantive interest, e. g., where a quantitative outcome variable can only be measured by qualitative observations such as solving or not solving certain items indicating the (latent) ability. However, it can also be of interest if the manifest measures are *linearly* related to the latent variable such as in models of classical test theory (see, e. g., Steyer, 2001) or in models of latent state-trait theory (see, e. g., ?, ?). If, e. g., there are three manifest variables Y_1, Y_2, and Y_3 measuring a single latent variable η, we may ask if there is just one single effect of the treatment on the latent outcome variable η — which transmits these effects to the manifest variables Y_1, Y_2, and Y_3 — instead of three separate effects of X on each variable Y_i. Hence, the latent variable may also be considered to be a mediator variable. Showing that all effects of X on the variables Y_i are indirect, i. e., mediated by η is one of the research efforts that aims at establishing construct validity of the latent variable η.

In the simplest case with a single latent variable, we consider the following single-unit trial:

(a) Sampling a person u out of a set of persons (the population of persons),
(b) assigning the unit or observing its selection to one of several experimental conditions (represented by the value x of the treatment variable X),

(d) recording the numerical values $y_1, \ldots, y_m$ of the manifest outcome variables $Y_1, \ldots, Y_m$.

In this single-unit trial, the values u of the observational-unit variable U again represent the observational *unit at the onset of treatment*, while the latent outcome variable η represents some attribute of the unit at the time point at which the outcome of the treatment is assessed. Clearly, this time point is *after* treatment and *prior* to the observation of the manifest outcome variables Y_i, at least as long as we preclude change in the latent variable during the process of assessing the manifest outcome variables. If this cannot be precluded, we would have to consider the time sequence in assessing the manifest outcome variables (e. g., of the items to be solved) as well.

Covariates

What are the covariates in such a single-unit trial? Again, the answer depends on the cause considered. If it is the treatment variable X, then each attribute of the *unit at the onset of treatment* is a covariate (with respect to X). Obviously, this again includes variables such as *sex*, *race*, and *educational status*. Note that in this kind of experiments, the set of covariates of X is the same irrespective of the choice of the outcome variable. Remember, we may not only consider the *latent* outcome variable η but also the *manifest* outcome variables Y_i, e. g., in order to study whether or not the effects of X on these manifest outcome variables are perfectly transmitted (or mediated) through the latent variable η.

Choosing the latent outcome variable η as a cause of the manifest outcomes variables Y_i brings additional covariates into play, for instance, all those variables that are *in between* treatment and the assessment of η. If, e. g., we consider an experiment studying the effects of different teaching methods, these additional covariates are critical life events (such as father or mother leaving the family), or additional lessons taken after treatment and before outcome assessment, for instance.

2.7 Summary and Conclusions

In this chapter we described a number of random experiments in informal terms. The purpose was to get a first idea which kind of empirical phenomena causal theories and hypotheses refer to. We focused on single-unit trials, which are the kinds of empirical phenomena we are interested in, both in theory and practice. We emphasized that a single-unit trial is a random experiment and discussed several kinds of random variables playing a crucial role in the theory of causal effects. We also mentioned that there is a certain *time order* among these random variables, e. g., saying that the covariates are 'prior' or 'simultaneous' to the treatment variable, which itself is 'prior' to the outcome variable. Furthermore, for each single-unit trial and each cause in such a single-unit trial, we discussed the

Box 2.1 Glossary of New Concepts

Random experiment	The kind of empirical phenomenon that events, random variables, and their dependencies refer to.
Single-unit trial	A particular kind of random experiment that consists of sampling a unit from a set of observational units and observing the values of one or more random variables related to this unit.
Cause	A random variable. Its effect on an outcome variable is considered.
Outcome variable	A random variable. Its dependency on a cause is considered.
Covariate of a cause	A random variable that can never be affected by the cause. It is prior or simultaneous to the cause. It might be correlated with the cause and the outcome variable.
Fallible covariate	A covariate that is assessed with measurement error.
Latent covariate	A covariate that is not directly observed. Instead it is defined by a set of manifest variables and a measurement model describing the dependencies of the manifest variables on the latent covariate.
Intermediate Variable	A variable that might mediate (transmit) the effect of the cause on the outcome variable. The cause is always prior to an intermediate variable and an intermediate variable is always prior to the outcome variable. An intermediate variable is not *necessarily* affected by the cause and it does not *necessarily* have an effect on the outcome variable.
Mediator	An intermediate variable on which X has a causal effect and which itself has a causal effect on the outcome variable Y.

Note that all these terms are still of an informal nature. Their mathematical treatment starts in chapter 3.

covariates involved. We emphasized that each cause considered in such a single-unit trial has its own set of covariates.

Other Single-Unit Trials

The single-unit trials discussed in this chapter are just a small selection of single-unit trials in which causal effects and causality of stochastic dependencies are of interest. We might also consider single-unit trials with latent covariates *and* la-

tent outcome variables *and* manifest and/or latent intermediate variables, but also single-unit trials with multiple mediation. Furthermore, we could also consider single-unit trials of growth curve models (see, e. g., Biesanz, Deeb-Sossa, Aubrecht, Bollen, & Curran, 2004; Bollen & Curran, 2006; Meredith & Tisak, 1990; Singer & Willett, 2003; Tisak & Tisak, 2000), latent change models (see, e. g., McArdle, 2001; Steyer, Eid, & Schwenkmezger, 1997; Steyer, 2005), or cross-lagged panel models (see, e. g., Kenny, 1975; Rogosa, 1980b; Watkins, Lei, & Canivez, 2007; Wolf, Chandler, & Spies, 1981). Causality is also an issue in uni- and multivariate time-series analysis as well as in stochastic processes with continuous time. However, in this book our examples will usually deal with experiments and quasi-experiments, including latent covariates and outcome variables as well as intermediate variables.

Outlook

Steyer and Nagel (in press-a) study how random experiments and the dependencies between events and random variables can be represented in terms of probability theory. In chapter 3 we extend the mathematical structure so that we can also meaningfully talk about time order between events and random variables and distinguish between *covariates* and intermediate variables. This will provide the mathematical framework in which causal effects can be meaningfully discussed.

2.8 Exercises

▷ **Exercise 2-1** Imagine that the probabilities of a crash for a flight with Airline A is ten times smaller than with Airline B. Which airline would you choose?

▷ **Exercise 2-2** Why does the theory of causal effects refer to single-unit trials?

▷ **Exercise 2-3** Why is it important to know which random experiment we are talking about?

▷ **Exercise 2-4** Which type of random experiment did we refer to in Simpson's paradox and in the nonorthogonal ANOVA example described in chapter 1?

▷ **Exercise 2-5** Why is it important to emphasize that, in simple experiments and quasi-experiments (see section 2.1), the observational-unit variable U represents the observational units *at the onset of treatment*?

▷ **Exercise 2-6** What is the basic idea of a covariate pertaining to a cause?

▷ **Exercise 2-7** Which kinds of causal effects can be considered in the simple experiment or quasi-experiment in which no *fallible* covariate and no intermediate variable is assessed?

▷ **Exercise 2-8** Which are the covariates pertaining to an intermediate variable if it is considered a cause of the outcome variable?

Solutions

▷ **Solution 2-1** Of course, B. Note that we apply these probabilities to the random experiment of flying *once* with A or B, even if these probabilities have been estimated in a sample.

▷ **Solution 2-2** Within such a single-unit trial, the various concepts of causal effects can be defined and we can study how to identify these causal effects from the parameters describing the joint distribution of the random variables considered. In such a single-unit trial, there usually is a clear time order which helps to disentangle the possible causal relationships between the random variables considered.

▷ **Solution 2-3** Different random experiments are different empirical phenomena. Although the names of the variables in different random experiments might be the same, the variables themselves are different entities, implying that the dependencies and effects between these variables might be different in different random experiments.

▷ **Solution 2-4** The type of random experiment we refer to in these examples is the single-unit trial of simple experiments and quasi-experiments described in section 2.1.

▷ **Solution 2-5** In the social sciences, units are often persons, and persons can change over time. If, in a simple experiment or quasi-experiment, a value u of U represents the observational unit sampled *at the onset of treatment*, each covariate will be a function of U. If, in contrast, U would represent the *observational unit at the assessment of a fallible covariate* (see section 2.2), which is some time prior to the onset of treatment, there can be other covariates in between assessment of the fallible covariate and the onset of treatment. We have to consider these additional covariates both in the definition of causal effects and in data analysis.

▷ **Solution 2-6** A covariate pertaining to a cause is a variable that is prior or simultaneous to the cause.

▷ **Solution 2-7** If the treatment has just two values, say *treatment* and *control*, there are different kinds of causal effects of the treatment variable on the outcome variable Y, such as the average total treatment effect, the conditional total treatment effects given a value of a covariate Z, and the *individual total effect* of X on Y given an observational unit u. Aside from these treatment effects, we may also consider the causal effects of a covariate Z on the treatment variable X, but also on the outcome variable Y. Among the causal effects of such a covariate are its conditional direct effects on Y given the different values of the treatment variable, the average of these X-conditional direct effects, and its indirect effect mediated by X.

▷ **Solution 2-8** Covariates pertaining to such an intermediate variable M are all variables representing attributes of the observational units at the onset of treatment, all variables that are simultaneous to treatment, including X itself, all other variables that are in between treatment and the intermediate variable.

References

Abraham, W. T., & Russell, D. W. (2004). Missing data: A review of current methods and applications in epidemiological research. *Current Opinion in Psychiatry*, *17*, 315–321.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Aiken, L. S., & West, S. G. (1996). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Aitkin, M. (1978). The analysis of unbalanced cross-classifications. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, *141*, 195–223.

Allen, M. P. (1997). *Understanding regression analysis*. New York, NY: Plenum Press.

Appelbaum, M. I., & Cramer, E. M. (1974). Some problems in the nonorthogonal analysis of variance. *Psychological Bulletin*, *81*, 335–343.

Arbuckle, J. L. (2006). AMOS 7.0 user's guide [Computer software manual]. Chicago, IL: SPSS.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Bauer, H. (1996). *Probability theory*. Berlin, Germany and New York, NY: de Gruyter.

Bentler, P. M. (1995). EQS Structural Equations program manual [Computer software manual]. Encino, CA: Multivariate Software.

Bentler, P. M., & Wu, E. J. C. (2002). *EQS 6 for Windows guide*. Encino, CA: Multivariate Software.

Berger, M. P. F., & Wong, W. K. (Eds.). (2005). *Applied optimal designs*. Chichester, England: Wiley.

Biesanz, J. C., Deeb-Sossa, N., Aubrecht, A. M., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, *9*, 30–52.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605–634.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach*. Hoboken, NJ: Wiley.

Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics*, *43*, 951–973.

Borooah, V. K. (2001). *Logit and probit: Ordered and multinomial models*. Thousand Oaks, CA: Sage.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203–219. doi: 10.1037/0033-295X.110.2.203

Browne, M. W., & Mels, G. (1998). Path analysis: RAMONA. In *SYSTAT for Windows: Advanced Applications (Version 8) [Computer software manual]*. Evanston, IL: SYSTAT.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago, IL: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Carlson, J. E., & Timm, N. H. (1974). Analysis of nonorthogonal fixed-effects designs. *Psychological Bulletin*, *81*, 563–570.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, *13*, 419–437.

Cheng, J., & Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *68*, 815–836.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, *13*, 261–281.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Cox, D. R., & Wermuth, N. (2004). Causality: A statistical view. *International Statistical Review*, *72*, 285–305.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *41*(1), 1–31.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. New York, NY: Wiley.

Dunn, G., Maracy, M., Dowrick, C., Ayuso-Mateos, J. L., Dalgard, O. S., Page, H., ... Wilkinson, G. (2003). Estimating psychological treatment effects from a randomised controlled trial with both non-compliance and loss to follow-up. *British Journal of Psychiatry*, *183*, 323–331.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on Generalized Linear Models* (2nd ed.). New York, NY: Springer.

Feller, W. (1968). *An introduction to probability theory and its applications* (2nd ed., Vol. 1). New York, NY: Wiley.

Feller, W. (1971). *An introduction to probability theory and its applications* (2nd ed., Vol. 2). New York, NY: Wiley.

Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know. *Organizational Research Methods*, *6*, 282–308.

Fisher, R. A. (1925/1946). *Statistical methods for research workers* (10th ed.). Edinburgh, England: Oliver and Boyd.

Gelman, A., & Hill, J. L. (2007). *Data analysis using regression and multilevel / hierarchical models.* New York, NY: Cambridge University.

Georgii, H.-O. (2008). *Stochastics – Introduction to probability and statistics.* Berlin, Germany: de Gruyter.

Gosslee, D. G., & Lucas, H. L. (1965). Analysis of variance of disproportionate data when interaction is present. *Biometrics, 21*, 115–133.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology, 78*, 119–128.

Green, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Greenland, S. (2000). Causal analysis in the health sciences. *Journal of the American Statistical Association, 95*, 286–289.

Greenland, S. (2004). An overview of methods for causal inference from observational studies. In A. Gelman & X.-L. Meng (Eds.), *Applied bayesian modeling and causal inference from incomplete-data perspectives* (pp. 3–14). Chichester, England: Wiley.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology, 10*, 37–48.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology, 15*, 413–419.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression.* New York, NY: Springer.

Hancock, G. R. (2004). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences.* Thousand Oaks, CA: Sage.

Hernán, M. A., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, 1–6. (Advance Access published March 30, 2011) doi: 10.1093/ije/dyr041

Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Research Methodology, 5*, 1–12.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. *Sociological Methodology, 18*, 449–484.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.

Hoyer, J., & Klein, A. (2000). Self-reflection and well-being: Is there a healthy amount of introspection? *Psychological Reports, 86*, 135–141.

Huet, S., Bouvier, A., Poursat, M.-A., & Jolivet, E. (2004). *Statistical tools for nonlinear regression: A practical guide with S-PLUS and R examples* (2nd ed.).

New York, NY: Springer.

Jamieson, J. (2004). Analysis of covariance ANCOVA with difference scores. *International Journal of Psychophysiology*, *52*, 277–283.

Jennings, E., & Green, J. L. (1984). Resolving nonorthogonal ANOVA disputes using cell means. *Journal of Experimental Education*, *52*, 159–162.

Jo, B. (2002a). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*, *27*, 385–409.

Jo, B. (2002b). Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. *Statistics in Medicine*, *21*, 3161–3181.

Jo, B. (2002c). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods*, *7*, 178–193.

Jo, B., Asparouhov, T., Muthén, B. O., Ialongo, N. S., & Brown, C. H. (2008). Cluster randomized trials with treatment noncomplicance. *Psychological Methods*, *13*, 1–18.

Jöreskog, K. G., & Sörbom, D. (1996/2001). LISREL 8: User's Reference Guide. [Computer software manual].

Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions*. Thousand Oaks, CA: Sage.

Keele, L. (2008). *Semiparametric regression for the social sciences*. Chichester, England: Wiley.

Kenny, D. A. (1975). Cross-lagged panel correlation: A test for spuriousness. *Psychology Bulletin*, *82*, 887–903.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*, 201–210.

Kenny, D. A., & Judd, C. M. (1996). A general procedure for the estimation of interdependence. *Psychological Bulletin*, *119*(1), 138–148.

Keren, G., & Lewis, C. (1976). Nonorthogonal designs: Sample versus population. *Psychological Bulletin*, *83*, 817–826.

King, G., & Zeng, L. (2001). Improving forecasts of state failure. *World Politics*, *53*, 623–658.

Klauer, K. J., Willmes, K., & Phye, G. D. (2002). Inducing inductive reasoning: Does it transfer to fluid intelligence? *Contemporary Educational Psychology*, *27*, 1–25.

Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457–474.

Klein, A. G., & Muthén, B. O. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, *42*, 647–673.

Klenke, A. (2013). *Probability theory – A comprehensive course* (2nd ed.). London, England: Springer. doi: 10.1007/978-1-4471-5361-0

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (2nd ed.; N. Morrison, Trans.). New York, NY: Chelsea.

Kramer, C. Y. (1955). On the analysis of variance of a two-way classification with unequal sub-class numbers. *Biometrics*, *11*, 441–452.

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, *13*, 163–167.

Lechner, M. (2000, November). A note on the "common support problem" in applied evaluation studies. Retrieved from `http://ssrn.com/abstract=259239`

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346. doi: 10.1002/sim.3782

Lee, S.-Y. (2007). *Structural equation modeling: A bayesian approach*. Chichester, England: Wiley.

Liao, T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models*. London, England: Sage.

Little, T. D., Card, N. A., Bovaird, J. A., Preacher, K. J., & Crandall, C. S. (2007). Structural equation modeling of mediation and moderation with contextual factors. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 207–230). Mahwah, NJ: Lawrence Erlbaum.

Loève, M. (1977). *Probability theory I* (4th ed.). New York, NY: Springer.

Loève, M. (1978). *Probability theory II* (4th ed.). New York, NY: Springer.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Lawrence Erlbaum.

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, *9*, 275–300.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2014). Theory and analysis of total, direct, and indirect causal effects. *Multivariate Behavioral Research*, *49*(5), 425–442.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 341–380). Lincolnwood, IL: Scientific Software International.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425.

McCullagh, P., & Nelder, J. A. (1989). *Monographs on statistics and applied probability: Vol. 37. Generalized linear models* (2nd ed.; D. R. Cox, D. V. Hinkley,

N. Reid, D. B. Rubin, & D. V. Silverman, Eds.). Chapman & Hall.

Meredith, M., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Mill, J. S. (1843/1865). Of the four methods of experimental inquiry. In *A system of logic, ratiocinative and inductive: Volume 1. Being a connected view of the principles of evidence, and the methods of scientific investigation.* London, England: Longmans, Green, and Co.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research*, *35*, 3–60. Retrieved from `http://smr.sagepub.com` doi: 10.1177/0049124106289164

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference. methods and principles for social research.* New York, NY: Cambridge University.

Muthén, L. K., & Muthén, B. O. (1998-2007). Mplus User's Guide (5th ed.) [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620.

Nagengast, B. (2009). *Causal inference in multilevel models* (Unpublished doctoral dissertation). Friedrich-Schiller-Universität Jena, Thüringen, Germany.

Nagengast, B., Kröhne, U., Bauer, M., & Steyer, R. (2007). Causal Effects Explorer: A didactic tool for teaching the theory of individual and average causal effects [Computer software manual]. University of Jena, Thüringen, Germany.

Nelder, J. A., & Lane, P. W. (1995). The computer analysis of factorial experiments: In memoriam – Frank Yates. *The American Statistician*, *49*, 382–385.

OpenMx. (2009). *OpenMx - Advanced Structural Equation Modeling [Computer Software].* Retrieved from http://openmx.psyc.virginia.edu/.

Overall, J. E., & Spiegel, D. K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin*, *72*(5), 311–322.

Overall, J. E., & Spiegel, D. K. (1973a). Comment on "Regression analysis of proportional cell data". *Psychological Bulletin*, *80*, 28–30.

Overall, J. E., & Spiegel, D. K. (1973b). Comments on Rawlings' nonorthogonal analysis of variance. *Psychological Bulletin*, *79*, 164–167.

Overall, J. E., Spiegel, D. K., & Cohen, J. (1975). Equivalence of orthogonal and nonorthogonal analysis of variance. *Psychological Bulletin*, *82*, 182–186.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, *27*, 226–284.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK.

Pearson, K., Lee, A., & Bramley-Moore, L. (1899). Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred race-

horses. *Series A, Containing Papers of a Mathematical or Physical Character: Philosophical Transactions of the Royal Society of London*, *192*, 257–330.

Pukelsheim, F. (2006). *Optimal design of experiments*. Philadelphia, PA: SIAM (Society for Industrial and Applied Mathematics).

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: Wiley.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* Princeton, NJ: Educational Testing Service: Educational Testing Service.

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. C. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings*. Orlando, FL: Houghton Mifflin.

Reichardt, C. S. (2005). Nonequivalent group design. In B. Everitt & D. Howell (Eds.), *Encyclopedia of behavioral statistics*. New York, NY: Wiley.

Robins, J., & Rotnitzky, A. (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, *91*, 763–783.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine*, *17*, 269–302.

Rogosa, D. (1980a). Comparing nonparallel regression lines. *Psychological Bulletin*, *88*, 307–321.

Rogosa, D. (1980b). A critique of cross-lagged correlation. *Psychological Bulletin*, *88*, 245–258.

Rosenbaum, P. R. (2002). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, *97*(457), 183–192. doi: 10.1198/016214502753479329

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *45*, 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. doi: 10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1984). Assessing the fit of logistic regressions using the implied discriminant-analysis: Comment. *Journal of the American Statistical Association*, *79*, 79–80.

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, *31*, 161–170.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331. doi: 10.1198/016214504000001880

Rubin, D. B. (2006). *Matched sampling for causal effects*. New York, NY: Cambridge University.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, *25*, 4334–4344.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *13*, 238–241.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, *33*, 230–251. doi: 10.3102/1076998607307239

Song, X. Y., & Lee, S. Y. (2006). Bayesian analysis of structural equation models with nonlinear covariates and latent variables. *Multivariate Behavioral Research*, *41*, 337–365.

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika*, *43*, 381–396.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT.

Splawa-Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (Reprinted from roczniki nauk rolniczych tom 10, pp. 1–51, 1923). *Statistical Science*, *5*, 465–480.

Spohn, W. (1980). Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, *9*, 73–99.

Stegmüller, W. (1983). *Erklärung, Begründung, Kausalität: Probleme und Resultate der Wissenschaftstheorie und analytischen Philosophie. [Explanation, justification, causality: Problems and findings of philosophy of science and analytic philosophy]*. Berlin, Germany: Springer.

Steyer, R. (1984). Causal linear stochastic dependencies: An introduction. In J. R. Nesselroade & A. von Eye (Eds.), *Individual development and social change: Explanatory analysis*. New York, NY: Academic Press.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability and testability. *Methodika*, *3*, 25–60.

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle [Theory of causal regression models]*. Stuttgart, Germany: Fischer.

Steyer, R. (2001). Classical test theory. In C. Ragin & T. D. Cook (Eds.), *International encyclopedia of the social and behavioural sciences: Logic of inquiry*

and research design (pp. 481–520). Oxford, England: Pergamon.

Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology, 1*, 39–54.

Steyer, R., & Eid, M. (2001). *Messen und Testen: Ein Lehrbuch [Measurement and testing]*. Berlin, Germany: Springer.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online, 2*, 21–33.

Steyer, R., Gabler, S., von Davier, A. A., & Nachtigall, C. (2000). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online, 5*, 55–87.

Steyer, R., & Nagel, W. (in press-a). *Probability and conditional expectation: Fundamentals for the empirical sciences*. Chichester, England: Wiley.

Steyer, R., & Nagel, W. (in press-b). *Probability and conditional expectation: Fundamentals for the empirical sciences*. Chichester, England: Wiley.

Steyer, R., & Partchev, I. (2007). EffectLite: User's manual – A program for the uni- and multivariate analysis of unconditional, conditional and average mean differences between groups [Computer software manual]. University of Jena, Thüringen, Germany.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam, Netherlands: North-Holland.

Tisak, J., & Tisak, M. S. (2000). Permanency and ephemerality of psychological measures with application to organizational commitment. *Psychological Methods, 5*, 175–198.

van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*, 920–925.

von Eye, A., & Schuster, C. (1998). *Regression analysis for social sciences*. San Diego, CA: Academic Press.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin, 109*, 147–151.

Wall, M. M., & Amemiya, Y. (2003). A method of moments technique for fitting interaction effects in structural equation models. *British Journal of Mathematical and Statistical Psychology, 56*, 47–63.

Watkins, M. W., Lei, P., & Canivez, G. L. (2007). Psychometric intelligence and achievement: A cross-lagged panel analysis. *Intelligence, 35*, 59–68.

West, S. G., & Aiken, L. S. (2005). Multiple linear regression. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Chichester, England: Wiley.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). New York, NY: Cambridge University.

Williams, J. D. (1972). Two way fixed effects analysis of variance with disproportionate cell frequencies. *Multivariate Behavioral Research, 7*, 57–83.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659–707.

Wolf, F. M., Chandler, T. A., & Spies, C. J. (1981). A crossed-lagged panel analysis of quality of school life and achievement responsibility. *Journal of Educational Research*, *74*, 363–368.

Woo, M.-J., Reiter, J. P., & Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in Medicine*, *27*, 3805–3816.

Woodward, J. A., & Bonett, G. B. (1991). Simple main effects in factorial designs. *Journal of Applied Statistics*, *18*(2), 255–264. doi: 10.1080/02664769100000019

Wright, S. (1918). On the nature of size factors. *Genetics*, *3*, 367–374.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*(7), 557–585.

Wright, S. (1923). The theory of path coefficients: A reply to Niles's criticism. *Genetics*, *8*, 239–255.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, *5*, 161–215.

Wright, S. (1960a). Path coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, *16*(2), 189–202.

Wright, S. (1960b). The treatment of reciprocal interaction, with or without lag, in path analysis. *Biometrics*, *16*(3), 423–445.

Yule, G. U. (1903). Notes on the theory of association of attributes of statistics. *Biometrika*, *2*, 121–134.