Methods of Psychological Research Online 2000, Vol.5, No.2
Internet: http://www.mpr-online.de Institute for Science Education

Causal Regression Models I: Individual and Average Causal Effects[1]

By Rolf Steyer[2]
Siegfried Gabler[3]
Alina A. von Davier[4]
Christof Nachtigall[4]
and
Thomas Buhl[4]

Abstract

We reformulate the theory of individual and average causal effects developed by Neyman, Rubin, Holland, Rosenbaum, Sobel, and others in terms of probability theory. We describe the kind of random experiment to which the theory refers, define individual and average causal effects, and study the relation between these concepts and the conditional expected value $E(Y \mid X = x)$ of the response Y in treatment condition x. For simplicity, we restrict our discussion to the case where there is no concomitant variable or covariate. We define the differences $E(Y \mid X = x_i) - E(Y \mid X = x_j)$ between these conditional expected values – the prima facie effects $[PFE(i, j)]$ – to be causally unbiased if the prima facie effect is equal to the average (of the individual) causal effects [$ACE(i$,

---

1 We would like to thank Albrecht Iseler (Free University of Berlin) for discussions in which the idea for some parts of this paper has been developed. We would also like to extend our gratitude to Angelika van der Linde (University of Edinburgh), to Donald B. Rubin (Harvard University), and to Peter Kischka (FSU Jena) who all gave helpful critical comments on previous versions of this paper. The first author also profited from extended discussions with Judea Pearl which helped to sharpen the theory presented. Thanks are also due to Katrin Krause.

2 Address for correspondence: Prof. Dr. Rolf Steyer, Friedrich Schiller University Jena, Am Steiger 3 - Haus 1, D-07743 Jena, Germany. Email: rolf.steyer@uni-jena.de

3 Center for Surveys, Methods, and Analyses (ZUMA), Mannheim, Germany

4 Friedrich-Schiller-University Jena, Germany

j)]. This equation, $PFE(i, j) = ACE(i, j)$, holds if the observational units are randomly assigned to the two experimental conditions. Thus, the theory justifies and gives us a deeper understanding of the randomized experiment. The theory is then illustrated by some examples. The first example demonstrates the crucial role of randomization, the second one shows that there are applications in which the observational units are not persons but persons-in-a-situation, and the third one demonstrates that causal unbiasedness of prima facie effects may be incidental. Specifically it is shown that although $PFE(i, j) = ACE(i, j)$ holds in the total population, the corresponding equations may not hold in any subpopulation. Hence, prima facie effects in the subpopulations might be seriously biased although they are causally unbiased in the total population. We argue that the theory has another serious limitation: A proposition that $PFE(i, j) = ACE(i, j)$ holds in the total population is not empirically falsifiable. Therefore there is a need for a more restrictive causality criterion other than causal unbiasedness that also has empirically testable implications.

Keywords: Causality; Confounding; Regression Models; Simpson Paradox; Experiment; Randomization; Rubin's Approach to Causality

Introduction

In many empirical studies we compare the means of a *response variable* Y between two or more treatments hoping that the difference between the means of Y in the two groups is an estimate of the causal effect of the *treatment variable* X on Y. In medicine we are interested, for instance, in the effect of a treatment on physical variables such as severity of headache, cancer vs. no cancer, or hormone concentration in the blood. In psychological treatments we might be interested in its effects on well-being, psychic health, or children's psychosocial behavior. In education we might like to know the effects of teaching or teaching styles on aptitudes, or knowledge. In all these studies X is a categorical variable with two or more values x_i and Y is a continuous real-valued variable or a dichotomous variable with values 0 or 1, and in all these studies the difference between the means of Y in the treatment groups may or may not convey information about the effect of X on the means of Y.[5]

[5] Note that the mean of Y is a relative frequency if Y is dichotomous with values 0 and 1.

In this paper we will introduce more precise concepts of causal effects that will help us to better understand experimental control techniques, such as randomization, and to select statistical techniques for the analysis of causal effects. Of course, these concepts are not entirely new. Instead they have been developed in the statistical literature almost during the whole 20th century. To our knowledge, Jerzy Neyman was the first who developed very concrete statistical concepts related to causality. Although, in this paper, we will focus the concepts going back to Neyman, it should be noted that there are other traditions related to 'statistics and causality' as well. Important recent contributions may be found under the keywords *graph theory*, for instance (see e.g., Pearl, 1995, 2000; Spirtes, Glymour & Scheines, 1993; Whittaker, 1990).[6]

Neyman (1923/1990, 1935) introduced the notion of an *individual causal effect*: the difference $E(Y \mid X = x_i, U = u) - E(Y \mid X = x_j, U = u)$ between the two conditional expected values of a response variable Y given an observational unit u and an experimental condition x_i and x_j, respectively. Unfortunately, in many applications, estimating both $E(Y \mid X = x_i, U = u)$ and $E(Y \mid X = x_j, U = u)$ for the *same* observational unit u would invoke untestable and often implausible auxiliary assumptions such as the nonexistence of position effects and transfer effects (see, e.g., Cook & Campbell, 1979; Holland, 1986). However, Neyman showed that the *average causal effect*, i.e., the average of the individual causal effects across the population of observational units, can be estimated by an estimate of the difference $E(Y \mid X = x_i) - E(Y \mid X = x_j)$ between the conditional expected values of Y in the experimental conditions x_i and x_j provided that the observational units are randomly assigned to the experimental conditions, and, of course, provided that this random assignment is not destroyed by systematic attrition (see, e.g., Cook & Campbell, 1979 or Rubin (1976).

Donald Rubin adopted Neyman's ideas, defined the individual causal effects in terms of observed scores (instead of expected values), developed his own notational system, and enriched this approach in a series of papers (e.g., Rubin, 1973a, b, 1974, 1977, 1978, 1985, 1986, 1990). His efforts have been continued by others such as Holland and Rubin (1983, 1988), Holland (1986, 1988), Rosenbaum and Rubin (1983a, b, 1984, 1985a, b), Rosenbaum (1984a, b, c), and Sobel (1994, 1995), for instance.

---

[6] See Pedhazur and Pedhazur Schmelkin (1991, Ch. 24), or Steyer, 1992, for a review of the literature on causality.

In this paper, we reformulate this theory in terms of classical probability theory. With this reformulation we do not intend to present a new theory or new concepts. The only purposes are to avoid problems in comprehending the original formulation, that have been addressed, e.g., by Dawid (1979, p. 30),[7] and to avoid the deterministic nature of the concepts in the original formulations of the theory. (For a critique, see, e.g., Dawid, 1997, pp. 10). Another goal of the paper is to provide some examples which illustrate the theory and provide the basis for discussing both, its merits and limitations. Some of these examples are analogous to the Simpson paradox (Simpson, 1951) which is well-known in categorical data analysis.

The paper is organized as follows: *First*, we describe the class of empirical phenomena to which the theory refers and introduce the notation and fundamental assumptions. *Second*, we reformulate the theory of individual and average causal effects in terms of probability theory. *Third*, we illustrate the theory by three examples. Finally, we discuss the merits and limitations of the theory, preparing the ground for a more complete theory of causal regression models (Steyer, Gabler, von Davier & Nachtigall, in press).

1. The Single-Unit Trial, Notation, and Assumptions

In classical probability theory, every stochastic model is based on at least one probability space consisting of:

(a) a set Ω of (possible) outcomes of the random experiment considered,

(b) a set $\mathfrak{A}$ of (possible) events, and

(c) a probability measure P assigning a (usually unknown) probability to each of the possible events.

In applications such a probability space represents the random experiment (the empirical phenomenon) considered.

---

[7] In the original formulation of the theory a different response variable is assumed for each treatment condition and the notation used presumes that there is a joint distribution of these different response variables. However, such a joint distribution is not quite easy to imagine because of the "fundamental problem of causal inference", namely that the same unit u cannot simultaneously be observed under each treatment condition. In section 3 we will show how this seeming contradiction in Rubin's theory can be resolved (see Footnote 12).

In this paper, the random experiment considered comprises the following components: Draw a unit u out of a set Ω_U (the population) of observational units (e.g., persons), assign the unit (or observe its self-assignment) to one of at least two experimental conditions gathered in the set Ω_X (e.g. , $\Omega_X = \{x_1, x_2\}$), and register the value $y \in \mathbb{R}$ of the response variable Y. This kind of random experiment may be referred to as the *single-unit trial*. Such a single-unit trial is already sufficient to define a regression or conditional expectation $E(Y|X)$ and discuss causality. It does not exclude other variables (e.g., mediators and covariates) to be observed as well.

The set of possible outcomes of the single-unit trial described above might be of the form:

$$\Omega = \Omega_U \times \Omega_X \times \mathbb{R}\,. \tag{1}$$

On such a set Ω, we can always construct an appropriate set $\mathfrak{A}$ of possible events, and $P(A)$ will denote the probability for each event $A \in \mathfrak{A}$.[8] The following random variables on such a probability space will be the other primitives of the theory of individual and average causal effects: The mapping U: $\Omega \to \Omega_U$, with $U(\omega) = u$, for each $\omega = (u, x, y) \in \Omega$, will denote the *observational-unit variable*, the mapping X: $\Omega \to \Omega_X$, with values x, will denote the *treatment variable*, and the function Y: $\Omega \to \mathbb{R}$, with values y, will represent the real-valued *response variable*.[9]

Note that the single-unit trial does not allow the mathematical analysis of parameter estimation and hypothesis testing. For these purposes we have to consider a *sampling model* consisting of a series of (usually independent) single-unit trials (see, e.g., Pratt & Schlaifer, 1988). However, remember that the definition of *stochastic dependence and independence* of two events A and B, for instance, is already meaningful for the random experiment of a *single* toss of two coins. Repeating such a random experiment *several* times is necessary only if we want to test if independence in fact holds or if we want to estimate the probabilities $P(A)$, $P(B)$, or $P(A \mid B)$, for instance. The concepts of stochastic dependence and independence, as well as the concept of probability apply to the random experiment of a *single* toss of two coins irrespective of whether or not this toss

---

8 $\mathfrak{A}$ is a σ-algebra, i.e., a set of subsets of Ω such that (a) $\Omega \in \mathfrak{A}$, (b) $A \in \mathfrak{A} \Rightarrow \overline{A} \in \mathfrak{A}$, and (c) for every sequence of elements of $\mathfrak{A}$, their union is also an element of $\mathfrak{A}$ (see, e.g. Bauer, 1981, p. 4).

9 Note that only Y has to be numerical. The general concepts of nonnumeric random variables and their distributions may be found in Bauer (1981), Dudley (1989), or Williams (1991), as well as the other concepts of probability theory, such as conditional expectations, for instance.

of two coins is embedded in a larger series of such coin tosses. Similarly, we can already define a regression $E(Y|X)$ and discuss the issue of causality for the type of random experiments described in this section.

In this paper, we will restrict our discussion to the single-unit trial. This will simplify notation and allow to focus attention on the central issues of this article: the relationship between individual and average causal effects on one side and the conditional expected values $E(Y|X=x)$ and their differences on the other side. Of course, considering a single-unit trial restricts the range of problems that can meaningfully be discussed. Especially, it will not be possible to discuss the problems of applying the treatments to several units simultaneously (cf. the discussion of SUTVA by Rubin, 1986). Another restriction is to not treat the more sophisticated case in which there is a concomitant variable or covariate Z (see, e.g., Sobel, 1994). The reason is to focus on the core of the theory in order to present it as simply and clearly as possible.

In the context presented in this paper, the regression $E(Y|X)$ may be called the *treatment* regression of Y, and $E(Y \mid X, U)$ the *unit-treatment regression* of Y. The residual $Y - E(Y \mid X, U)$ may contain different kinds of error components.[10] The notation and the assumptions will now be summarized in the following definition.

Definition 1. *$\langle (\Omega, \mathfrak{A}, P), E(Y|X), U \rangle$ is called a potential causal regression model with a discrete observational-unit variable and a discrete treatment variable if:*

(*a*) $(\Omega, \mathfrak{A}, P)$ *is a probability space;*

(*b*) $U: \Omega \rightarrow \Omega_U$*, the observational-unit variable, is a random variable on* $(\Omega, \mathfrak{A}, P)$*, where* Ω_U *is the set of "observational units", the "population";*

(*c*) $X: \Omega \rightarrow \Omega_X$*, the treatment variable, is a random variable on* $(\Omega, \mathfrak{A}, P)$*, where* Ω_X *is a finite set of "treatment conditions";*

[10] Neyman (1935) used the term "technical errors" in this context. Hence, our theory and notation does not presume that X has a deterministic causal effect, even not on the individual level. This indeterminism may be due to measurement error and/or to mediating variables and processes. The individual effect of treating Fritz with a psychotherapy against not treating him (control) would not be observable even if we could do both, treat him and not treat him. The reason is that there are many mediating variables such as critical (death of a dear friend) or fortunate (meeting the love of his life) life events that will change his observable response. (For another example, see Dawid, 1997, p. 10.) Hence, defining the individual effect via conditional expected values seems to be more realistic. Note that the deterministic case is included as a special case.

(*d*) $Y: \Omega \rightarrow \mathbb{R}$, *the response variable, is a real-valued random variable on* $(\Omega, \mathfrak{A}, P)$ *with positive and finite variance;*

(*e*) $E(Y|X)$ *is the regression (i.e., the conditional expectation) of Y on X;*

(*f*) $P(X = x, U = u) > 0$ *for each pair* (x, u) *of values of X and U.*

Assuming that the mappings U, X, and Y are random variables on the same probability space means to assume that they have a joint distribution. Assuming Y to have a finite variance $Var(Y)$ implies that the regressions (or, synonymously, conditional expectations) $E(Y|X)$ and $E(Y|X, U)$ exist. This variance might be determined by the regression $E(Y|X)$ of Y on X to some degree represented by the coefficient of determination $Var[E(Y|X)] \,/\, Var(Y)$. Assuming $P(X = x, U = u) > 0$ for each pair (x, u) of values of X and U implies that the individual conditional expected values $E(Y|X = x, U = u)$ of Y given $X = x$ and $U = u$ (see, e.g., Dawid, 1979; Neyman, 1923/1990), the values of the regression $E(Y|X, U)$, are uniquely defined. From a substantive point of view, this means that we will only be able to deal with *discrete* treatment variables X and *discrete* unit variables U.[11]

A *potential* causal regression model with discrete units and discrete treatment variable is a framework in which the issue of causality can be discussed. An *actual* causal regression model will additionally consist of a *causality criterion* that may or may not hold in an empirical application. Such causality criteria will be discussed later on and in papers to follow.

Whereas most points in the definition above are more or less formal requirements, there are some terms which have to be interpreted in empirical applications. For instance, we have to know the set Ω of possible outcomes of the random experiment considered and its components: the population of observational units Ω_U, the set of treatment conditions Ω_X, and the measurement procedure determining the score y of the response variable Y. Furthermore, we have to know, which is the observational-unit variable U, which is the treatment variable X, which is the response variable Y, and which is the regression $E(Y|X)$ that possibly describes a causal dependence. In the subsequent sections, we will always presume that there is a potential causal regression model with discrete units and discrete treatment variable as defined above.

[11] If we would *not* introduce this assumption, then there might be a pair (x, u), with zero probability, for which the conditional expected value $E(Y | X = x, U = u)$ were not uniquely defined. For a more general theory, which can also handle continuous variables X and U see Steyer (1992).

2. Individual and Average Causal Effects: Basic Concepts

Substantive scientists are not *primarily* interested in group means or other aggregated statistics. Their primary interest is rather in the individual causal effects, the definition of which goes back at least to Neyman (1923/1990; 1935).

Definition 2. *The individual causal effect of x_i vs. x_j on (the expected value of) Y for unit u is the difference* [12]

$$ICE_u(i, j) := E(Y | X = x_i, U = u) - E(Y | X = x_j, U = u). \quad (2)$$

This definition does not refer to data but to a random experiment and its laws from the *pre facto* perspective, i.e., from the perspective *before* the random experiment is conducted.[13] Hence, we can consider both conditional expected values, $E(Y | X = x_i, U = u)$ and $E(Y | X = x_j, U = u)$, although, in practice, the unit can often be assigned to one single experimental condition, only. For instance, a novice cannot be simultaneously taught mathematics by a new teaching method, represented by $X = x_i$, *and* by a traditional method, represented by $X = x_j$; after the first treatment he will not be a novice any more. Therefore, it is often impossible to estimate the values of the difference (2) *individually* for each unit (novice) u. Hence, in these cases we can either estimate the conditional expected value $E(Y | X = x_i, U = u)$ *or* $E(Y | X = x_j, U = u)$, but not both. This has been called the *fundamental problem of causal inference* by Holland (1986).

The solution to this problem is well-known at least since Neyman (1935): estimating the average of the individual causal effects, the *average causal effect* (see also Rubin, 1974; Holland, 1986; Neyman, 1923/1990, p. 470). This can be achieved in a randomized experiment, for instance, as will be shown in the next paragraphs.

---

[12] Rubin defined potential response variables Y_i for each treatment condition x_i. In our notation we could define these variables for each value x_i by $Y_i(u) := E(Y | X = x_i, U = u)$, for all values u of U. Defined in this way, it is obvious that the variables Y_i, $i = 1, \ldots, n_x$ are functions of U and that they have a joint distribution with X. This implies that the variables $Y_1, \ldots, Y_{n_x}$ and X are stochastically independent (denoted $Y_1, \ldots, Y_{n_x} \perp X$) if U and X are stochastically independent. The condition $Y_1, \ldots, Y_{n_x} \perp X$ and $0 < P(X = x_i) < 1$, for each value x_i of X has been called "strong ignorability" (see, e.g., Rosenbaum & Rubin, 1983a, p. 213).

[13] In this context, other authors use the term "counterfactual" (see, e.g., Sobel, 1994). However, strictly speaking, this term implies a *post facto* perspective, which is not the perspective taken in a stochastic model.

In the following definition, the summation is across all values $u \in \Omega_U$, and $P(U = u)$ denotes the probability that the observational unit u is drawn.[14]

Definition 3. *The average causal effect of x_i vs. x_j on (the expected value of) Y is defined:*

$$ACE(i, j) := \sum_u ICE_u(i, j)\ P(U = u). \tag{3}$$

Note that $ACE(i, j)$ is not estimable, too, unless an assumption is introduced. In Section 4, we will introduce such an assumption that implies the equality of $ACE(i, j)$ and the *prima facie effects* which are defined as follows:

Definition 4. *The prima facie effect of x_i vs. x_j on (the expected value of) Y is defined:*

$$PFE(i, j) := E(Y | X = x_i) - E(Y | X = x_j). \tag{4}$$

It is the equality of the prima facie effect $PFE(i, j)$ and the average causal effect $ACE(i, j)$ which is meant when we say that the prima facie effect is *causally unbiased*. Hence, it will be useful to formally introduce this and some related concepts.

Definition 5. (i) *The term*

$$CUE(Y | X = x) := \sum_u E(Y | X = x, U = u)\ P(U = u) \tag{5}$$

is called the (causally) unbiased expected value of Y given x.

(*ii*) *A conditional expected value $E(Y | X = x)$ is defined to be causally unbiased if*

$$E(Y | X = x) = CUE(Y | X = x) \tag{6}$$

(*iii*) *The prima facie effect of x_i vs. x_j on (the expected value of) Y is defined to be causally unbiased if*

$$PFE(i, j) = ACE(i, j). \tag{7}$$

In Equation (5), again the summation is across all elements u of the population Ω_U (see Footnote 14). Note that, in Definition 5 we use the term "unbiased" although we

[14] In most cases, $P(U = u) = 1 / N$, where N denotes the number of observational units in the population Ω_U. Note, however, that the set Ω_U may also be interpreted as a set of persons in situations, for instance (see Example II). In these nonstandard applications the distributions of U would be unknown.

are *not* dealing with problems of statistical estimation. Specifically, the causally unbiased expected value of the response variable Y is "unbiased" or "unaffected" by the mechanism assigning a unit to the experimental condition x. This is not true for the conditional expected value $E(Y \mid X = x)$ [see Eq. (9)], because its computation indirectly involves the *conditional* probabilities $P(X = x \mid U = u)$ (*the individual assignment probabilities*) via the conditional probabilities $P(U = u \mid X = x) = P(X = x \mid U = u) \cdot P(U = u) \,/\, P(X = x)$.

Corollary 1. *If the conditional expected values* $E(Y \mid X = x_i)$ *and* $E(Y \mid X = x_j)$ *are causally unbiased, then the prima facie effect* $E(Y \mid X = x_i) - E(Y \mid X = x_j)$ *is causally unbiased, i.e., then* $PFE(i, j) = ACE(i, j)$.

This corollary is an immediate consequence of Definition 5. It emphasizes that the crucial point in the theory of individual and average causal effects is the distinction between the conditional expected value $E(Y \mid X = x)$ and the unbiased expected value $CUE(Y \mid X = x)$ of Y given $X = x$.

Note that the parameters $CUE(Y \mid X = x)$ are of substantive interest (see, e.g., Pratt & Schlaifer, 1988). In fact, aside from the individual causal effects it is the causally unbiased conditional expected values and their differences

$$ACE(i, j) = CUE(Y \mid X = x_i) - CUE(Y \mid X = x_j), \tag{8}$$

the average causal effects, that are of substantive interest. Of course, the concept of causal unbiasedness may be extended from these differences to other linear combinations of the conditional expected values $E(Y \mid X = x)$, other than the differences $PFE(i, j) = E(Y \mid X = x_i) - E(Y \mid X = x_j)$.

Note that the concept of causality associated with causally unbiased expected values and average causal effects is a rather weak one, because it only deals with the *average* of the individual causal effects [see Eq. (3)]. Hence, even if the average causal effect is positive, there can be observational units and subpopulations for which the individual or average causal effects are negative. Investigating questions like these is the focus of analyzing interactions in the analysis of variance or of moderator models in regression analyses. A stronger concept would require invariance of the individual effects (see, e.g., Steyer, 1985, 1992, Ch. 9) or at least invariance of the individual effects in subpopulations (Steyer, 1992, Ch. 14).

According to Equation (6) causal unbiasedness of a conditional expected value $E(Y \mid X = x)$ means that $E(Y \mid X = x)$ is equal to $CUE(Y \mid X = x)$, the (causally) unbiased expected value of Y. As mentioned before, computing the *conditional expected value* $E(Y \mid X = x)$ involves the *conditional probabilities* of $P(U = u \mid X = x)$. This can be seen from the equation

$$E(Y \mid X = x) = \sum_u E(Y \mid X = x,\ U = u)\ P(U = u \mid X = x), \tag{9}$$

which is always true provided that U is discrete. The conditional probabilities $P(U = u \mid X = x)$ are closely related to the treatment assignment probabilities $P(X = x \mid U = u)$. The latter directly reflect that some units may have a higher chance to be treated than others. This is the mechanism that creates causal bias of the conditional expected values $E(Y \mid X = x)$, at least in populations that are not homogeneous with respect to their conditional expected values $E(Y \mid X = x,\ U = u)$ (see Theorem 2 below). The computation of the unbiased expected value $CUE(Y \mid X = x)$ of Y given x only involves the *unconditional probabilities* $P(U = u)$ [see Eq. (5)].[15] Figure 1 summarizes the basic concepts introduced above.

[15] We choose to develop the theory of individual and average causal effects in a framework in which there is a nontrivial joint distribution of U and X. An alternative is to drop this prerequisite and the assumption that X has a distribution. We would then use the distributions of Y and U *within* fixed conditions x. From a formal point of view we would then need as many different probability spaces as there are different values x representing the treatment conditions. Other differences would only be in the notations.

Treatment variable	X (with values x, x_i, and x_j
Response variable	Y (with values $y \in$ IR)
Observational-unit variable	U (with values u, the "observational unit")
Individual conditional expected values	$E(Y \mid X = x, U = u)$
Individual causal effect	$ICE_u(i, j) = E(Y \mid X = x_i, U = u) - E(Y \mid X = x_j, U = u)$
Individual sampling probabilities	$P(U = u)$
Unbiased expected value of Y given x	$CUE(Y \mid X = x) = \sum_u E(Y \mid X = x, U = u)\, P(U = u)$
Average causal effect:	$ACE(i, j) = \sum_u ICE_u(i, j)\, P(U = u)$
Individual assignment probabilities	$P(X = x \mid U = u)$
Conditional expected values	$E(Y \mid X = x) = \sum_u E(Y \mid X = x, U = u)\, P(U = u \mid X = x)$
Prima facie effect	$PFE(i, j) = E(Y \mid X = x_i) - E(Y \mid X = x_j)$
Causal unbiasedness of $E(Y \mid X = x)$	$E(Y \mid X = x) = CUE(Y \mid X = x)$
Causal unbiasedness of the prima facie effect	$PFE(i, j) = ACE(i, j)$

Figure 1. *Basic concepts of the theory of individual and average causal effects*

3. Individual and Average Causal Effects: Theorems

Theorem 1 provides the theoretical foundation for a practical solution to the fundamental problem of causal inference. It provides the link between the (empirically estimable) conditional expected values and the prima facie effect on one side and the average of the individual conditional expected values and the average causal effect (the parameters of theoretical interest) on the other side.[16]

Theorem 1. [Stochastic independence of X and U]. *If U and X are stochastically independent, then each conditional expected value $E(Y \mid X = x)$ is causally unbiased.*

Stochastic independence of the observational-unit variable U and the treatment variable X can deliberately be created by the technique of random assignment of units to experimental conditions. Hence, Theorem 1 helps to understand the importance of this

16 The proofs will be found in Appendix A.

technique of experimentation. Together with Corollary 1 it proves that randomization is sufficient for causal unbiasedness of a prima facie effect.[17]

However, note that stochastic independence of U and X is *sufficient* but not *necessary* to imply causal unbiasedness. Here is a second sufficient condition that implies causal unbiasedness.

Theorem 2. [Unit-treatment homogeneity]. *If Y is X-conditionally regressively independent of U, i.e., if*

$$E(Y|X, U) = E(Y|X), \tag{10}$$

then each conditional expected value $E(Y|X = x)$ is causally unbiased.

Equation (10) may be called *unit-treatment homogeneity.* It means that under a given treatment condition x all units u have the same expected value $E(Y|X = x, U = u) = E(Y|X = x)$. Note that, in contrast to independence of U and X, Equation (10) is *not* under the experimenter's control. Although it will seldom hold, it nevertheless *is* a sufficient condition for causal unbiasedness which may be useful in *some* applications.

4. Example I

We will now present an example that may help to understand the basic concepts summarized in Figure 1, especially the distinction between $E(Y|X = x)$ and $CUE(Y|X = x)$. Table 1 displays three examples which differ from each other by different *individual treatment assignment probabilities*, i.e., the probabilities with which each unit is assigned to the treatment condition x_i (see the last three columns). In Example Ia, the individual treatment assignment probability depends on the person and its expected values under the treatment condition, whereas in Examples Ib and Ic the individual treatment assignment probabilities are the same for all units. Hence, in these examples there is stochastic independence of U and X. Example Ic illustrates that stochastic independence of U and X does not necessarily mean that the assignment probabilities

[17] This is true although there will be (small) dependencies between groups of units and the treatment variable in samples (due to sampling error). Dependencies greater than explainable by sampling error may be due to unsuccessful randomization (e.g., systematic attrition of subjects). However, in such a case we should talk about an *attempted* randomized experiment, not about a (*true*) randomized experiment.

are equal for each treatment x_i and x_j. Example Ib presents a design with equal cell probabilities and Example Ic a design with proportional cell probabilities. In sampling models this would correspond to equal and proportional cell frequencies, respectively. Table 2 gives an alternative presentation of these examples which may be more familiar to the reader.

We will now numerically identify the *substantive parameters* and then the *technical parameters* in this example. With "substantive parameters" we mean the terms that are defined within the theory of individual and average causal effects. These terms are unknown in ordinary empirical applications. The "technical parameters" are those that are defined already in ordinary stochastic models and can be estimated without assumptions such as independence of U and X or unit-treatment homogeneity (see Theorems 1 and 2).

Table 1. Three examples with different individual treatment assignment probabilities

Observational units	$P(U=u)$	$E(Y\|X=x_1, U=u)$	$E(Y\|X=x_2, U=u)$	Individual treatment assignment probabilities $P(X=x_1 \mid U=u)$		
				Example Ia	Example Ib	Example Ic
u_1	1/2	85	91	1/4	1/2	1/3
u_2	1/2	105	109	3/4	1/2	1/3

Note. According to the theorem of total probability, the unconditional probability for treatment assignment is $P(X = x_1) = \Sigma_{i=1}^{2} P(X=x_1 \mid U=u_i) \cdot P(U=u_i) = 1/2$ for Examples Ia and Ib and $P(X=x_1) = 1/3$ for Example Ic.

The Substantive Parameters. What is of utmost interest for substantive researchers are the individual expected values $E(Y \mid X=x,\ U=u)$ displayed in the columns 3 and 4 of Table 1 and in the cells of Table 2 as well as the *individual causal effects*

$$ICE_{u_1}(1,2) = E(Y \mid X=x_1,\ U=u_1) - E(Y \mid X=x_2,\ U=u_1) = 85 - 91 = -6$$

and

$$ICE_{u_2}(1,2) = E(Y \mid X=x_1,\ U=u_2) - E(Y \mid X=x_2,\ U=u_2) = 105 - 109 = -4.$$

Table 2. An alternative presentation of the examples in Table 1

Observational units	$P(U = u)$	Treatment	Example Ia		Example Ib		Example Ic	
			x_1	x_2	x_1	x_2	x_1	x_2
u_1	1/2		85 (1/8)	91 (3/8)	85 (1/4)	91 (1/4)	85 (1/6)	91 (2/6)
u_2	1/2		105 (3/8)	109 (1/8)	105 (1/4)	109 (1/4)	105 (1/6)	109 (2/6)

Note. The ratios in parentheses are the joint probabilities $P(U = u, X = x)$ of drawing unit u and giving it treatment x. These probabilities may be computed by $P(U = u, X = x) = P(X = x \mid U = u) \cdot P(U = u)$.

If these numbers were known we would have perfect causal knowledge about the effect of X:[18] Assigning unit u_1 to x_1 would yield the expected Y-value 85, whereas assigning unit u_1 to x_2 would yield the expected Y-value 91.

If these numbers are *not* known it is still informative to know the unbiased expected value of Y given x in the population, i.e.:

$$CUE(Y \mid X = x_1) = \sum_u E(Y \mid X = x_1, U = u)\, P(U = u) = 85 \cdot 1/2 + 105 \cdot 1/2 = 95$$

and

$$CUE(Y \mid X = x_2) = \sum_u E(Y \mid X = x_2, U = u)\, P(U = u) = 91 \cdot 1/2 + 109 \cdot 1/2 = 100,$$

as well as the difference between these two averages, the *average causal effect*,

$$ACE(1,2) = CUE(Y \mid X = x_1) - CUE(Y \mid X = x_2) = 95 - 100 = -5.$$

For a randomly drawn unit, the expected effect of doing x_1 instead of x_2 is −5. This is the best guess for the individual causal effect if no information whatsoever about the individual is available.

The Technical Parameters. The difficulty in causal inference is that the parameters computed above can be obtained without bias only in very special cases (see Theorems 1 and 2, for instance). To illustrate this difficulty by the example displayed in Table 1,

[18] Of course, we would only know the individual conditional expectations and the individual causal effects, not the actual scores of the response Y itself, which may also be affected by other causes and by measurement error.

note that assuming unequal individual assignment probabilities (Example Ia) yields causally *biased* conditional expected values:

$$E(Y \mid X = x_1) = \sum_u E(Y \mid X = x_1,\ U = u)\ P(U = u \mid X = x_1)$$
$$= 85 \cdot 1/4 + 105 \cdot 3/4 = 100,$$

$$E(Y \mid X = x_2) = \sum_u E(Y \mid X = x_2,\ U = u)\ P(U = u \mid X = x_2)$$
$$= 91 \cdot 3/4 + 109 \cdot 1/4 = 95.5.$$

By incidence, the conditional probabilities $P(U = u \mid X = x)$ are identical with the conditional probabilities $P(X = x \mid U = u)$ in this example. This follows from $P(U = u \mid X = x) = P(X = x \mid U = u) \cdot P(U = u) / P(X = x)$ and $P(X = x) = P(U = u) = 1/2$. Obviously, $E(Y \mid X = x_1) \neq CUE(Y \mid X = x_1)$ and $E(Y \mid X = x_2) \neq CUE(Y \mid X = x_2)$, and $PFE(1,2) = 100 - 95.5 = 4.5 \neq ACE(1,2)$.

In fact, in Example Ia, the prima facie effect is positive, namely +4.5, whereas the average causal effect is negative, namely −5 (see Table 3 for a summary).[19]

In contrast to Example Ia, assuming equal individual assignment probabilities (Examples Ib and Ic), i.e., stochastic independence of U and X, yields causally unbiased conditional expected values. For Examples Ib and Ic we get:

$$E(Y \mid X = x_1) = 85 \cdot 1/2 + 105 \cdot 1/2 = 95$$

and

$$E(Y \mid X = x_2) = 91 \cdot 1/2 + 109 \cdot 1/2 = 100.$$

In these two examples, the conditional probabilities $P(U = u \mid X = x)$ are *not* identical with the conditional probabilities $P(X = x \mid U = u)$. In these two examples, the formula $P(U = u \mid X = x) = P(X = x \mid U = u) \cdot P(U = u) / P(X = x)$ always yields 1/2, no matter which value of X and U we consider. For x_1 and u_1 in Example 1b, for instance, we receive

$P(U = u_1 \mid X = x_1) = P(X = x_1 \mid U = u_1) \cdot P(U = u_1) / P(X = x_1) = (1/2) \cdot (1/2) / (1/2) = 1/2$,

[19] The phenomenon that the difference between means in a total sample is positive whereas it is negative in *each* subsample is known as the Simpson paradox (see, e.g., Simpson, 1951; Yule, 1903).

and in Example 1c this equation yields the same result, although via other probabilities:

$P(U = u_1 | X = x_1) = P(X = x_1 | U = u_1) \cdot P(U = u_1) / P(X = x_1) = (1/3) \cdot (1/2) / (1/3) = 1/2$.

Table 3. Results for Examples Ia to Ic

Example Ia						
Observational units	x_1		x_2		$ICE_u(1,2)$	
u_1	85	(1/8)	91	(3/8)	−6	
u_2	105	(3/8)	109	(1/8)	−4	
$E(Y\|X=x)$	100		95.5		4.5	$PFE(1,2)$
$CUE(Y\|X=x)$	95		100		−5	$ACE(1,2)$

Example Ib						
Observational units	x_1		x_2		$ICE_u(1,2)$	
u_1	85	(1/4)	91	(1/4)	−6	
u_2	105	(1/4)	109	(1/4)	−4	
$E(Y\|X=x)$	95		100		−5	$PFE(1,2)$
$CUE(Y\|X=x)$	95		100		−5	$ACE(1,2)$

Example Ic						
Observational units	x_1		x_2		$ICE_u(1,2)$	
u_1	85	(1/6)	91	(2/6)	−6	
u_2	105	(1/6)	109	(2/6)	−4	
$E(Y\|X=x)$	95		100		−5	$PFE(1,2)$
$CUE(Y\|X=x)$	95		100		−5	$ACE(1,2)$

Note. The ratios in parentheses are the joint probabilities $P(U = u, X = x)$ of drawing unit u and giving it treatment x.

In these cases $E(Y | X = x_1) = CUE(Y | X = x_1)$ and $E(Y | X = x_2) = CUE(Y | X = x_2)$. Hence, with the individual assignment probabilities of Examples Ib and Ic, the prima facie effect $PFE(1,2) = E(Y | X = x_1) - E(Y | X = x_2) = 95 - 100 = -5$ *is* causally unbiased, i.e., the prima facie effect is equal to the average causal effect $ACE(1,2)$ (see again Table 3 for a summary).

Obviously, the phenomenon described above [$PFE(i,j) \neq ACE(i,j)$] is not possible, if the conditional probabilities $P(X = x_1 | U = u)$ of being assigned to the treatment condition x_1 are the same for each unit, i.e., if U and X are stochastically independent (see Examples Ib and Ic). Table 4 displays a summary of the relevant conditional and unconditional probabilities. In Example Ia, however, the assignment of the unit to the treatment follows the rule that the person with the smaller expected Y-value, has the lower probability of being assigned to x_1. In this case, U and X are *not* stochastically independent.[20] Hence, unless there is unit-treatment homogeneity [see Eq. (10)], the way in which the unit is assigned to the treatment conditions is crucial for causal unbiasedness of the prima facie effects.

Table 4. Probabilities used in Example I

		Example Ia				Example Ib				Example Ic			
		x_1		x_2		x_1		x_2		x_1		x_2	
Observational units	$P(U=u)$	$P(X=x_1 \mid U=u)$	$P(U=u \mid X=x_1)$	$P(X=x_2 \mid U=u)$	$P(U=u \mid X=x_2)$	$P(X=x_1 \mid U=u)$	$P(U=u \mid X=x_1)$	$P(X=x_2 \mid U=u)$	$P(U=u \mid X=x_2)$	$P(X=x_1 \mid U=u)$	$P(U=u \mid X=x_1)$	$P(X=x_2 \mid U=u)$	$P(U=u \mid X=x_2)$
u_1	1/2	1/4	1/4	3/4	3/4	1/2	1/2	1/2	1/2	1/3	1/2	2/3	1/2
u_2	1/2	3/4	3/4	1/4	1/4	1/2	1/2	1/2	1/2	1/3	1/2	2/3	1/2
	$P(X=x)$	1/2		1/2		1/2		1/2		1/3		2/3	

To summarize: We may distinguish between *substantive* parameters and *technical* parameters. In this example, the parameters of substantive interest are the individual expected values, the unbiased expected values $CUE(Y \mid X = x)$ of Y given x, and the differences between the individual expected values and their averages.[21] These parameters are the basic concepts of the theory of individual and average causal effects. Computing the average of the individual expected values involves the distribution of the observational units described by the *unconditional* probabilities $P(U = u)$. In empirical applications, the parameters of substantive interest can be estimated only under special circumstances. Two such circumstances are unit-treatment homogeneity and stochastic inde-

[20] In Footnote 12 we defined the potential outcome variables Y_1 and Y_2. In this example, the values of Y_1 are the numbers 85 (for u_1) and 105 (for u_2), and the values of Y_2 are the numbers 91 (for u_1) and 109 (for u_2) (see columns 1 and 2 in Table 1). In Example Ib and Ic there is also independence of X and the vector (Y_1, Y_2).

[21] Pratt and Schlaifer (1988) use the term "laws" in this context.

pendence of U and X. Stochastic independence of U and X means in this example that the individual treatment assignment probabilities $P(X = x \mid U = u)$ are constant across the observational units u.[22] These individual treatment assignment probabilities may be called *technical parameters*, because they are not of primary substantive interest. Nevertheless, it is these technical parameters that usually decide about causal unbiasedness of the conditional expected values $E(Y \mid X = x)$ and their differences, the prima facie effects.

5. Example II

We now present an example in which we do not assume $P(U = u) = 1/N$ anymore. The substantive theoretical background is based on the conception that, in psychology, the observational units are not persons but persons-in-a-situation (Anastasi, 1983; Steyer, Ferring & Schmitt, 1992; Steyer, Schmitt & Eid, 1999). This means that we will

Table 5. *Three examples with different individual treatment assignment probabilities*

Persons	*Situations*	*Observational units*	$P(U = u)$	$E(Y \mid X = x_1, U = u)$	$E(Y \mid X = x_2, U = u)$	Individual treatment assignment probabilities $P(X=x_1 \mid U = u)$		
						Example IIa	Example IIb	Example IIc
p_1	s_1	u_1	1/10	85	91	2/10	1/2	1/3
p_1	s_2	u_2	4/10	74	78	1/10	1/2	1/3
p_2	s_1	u_3	1/10	106	112	9/10	1/2	1/3
p_2	s_2	u_4	4/10	95	99	8/10	1/2	1/3

Note. According to the theorem of the total probability, the unconditional probability of treatment assignment is $P(X = x_1) = \Sigma_{i=1}^{4} P(X=x_1 \mid U = u_i) \cdot P(U = u_i) = .47$ for Example IIa, $P(X = x_1) = 1/2$ for Example IIb, and $P(X = x_1) = 1/3$ for Example IIc.

have different individual expected values of the response variable Y for the same person in different situations. As a substantive example consider an aptitude test Y, and the two situations s_1 and s_2 "*at least* vs. *less than four hours sleep* in the night before the

[22] In a subsequent paper (Steyer et al., 2000), we will show that there is a less restrictive sufficient condition for causal unbiasedness.

test is taken". Furthermore, assume both persons p_1 and p_2 are often out at night and hence in situation s_2 most of the time. What will be the average causal effect of a treatment x_1 vs. a treatment x_2? Clearly, the four combinations of persons and situations, i.e., the four units, do not have equal probabilities 1/4 anymore. This affects the average causal effect.

Table 5 displays the relevant parameters. Again, we treat three examples which differ from each other by assuming different probabilities with which the units are assigned to the treatment condition x_1 (see the last three columns).

The Substantive Parameters. What is of utmost interest for substantive researchers are again the individual expected values $E(Y \mid X = x, U = u)$ as well as the individual causal effects

$$ICE_{u_1}(1,2) = E(Y \mid X = x_1, U = u_1) - E(Y \mid X = x_2, U = u_1) = 85 - 91 = -6 ,$$

$$ICE_{u_2}(1,2) = E(Y \mid X = x_1, U = u_2) - E(Y \mid X = x_2, U = u_2) = 74 - 78 = -4 ,$$

$$ICE_{u_3}(1,2) = E(Y \mid X = x_1, U = u_3) - E(Y \mid X = x_2, U = u_3) = 106 - 112 = -6 ,$$

and

$$ICE_{u_4}(1,2) = E(Y \mid X = x_1, U = u_4) - E(Y \mid X = x_2, U = u_4) = 95 - 99 = -4.$$

Again, if these numbers (as well as the persons and the situations in which the persons are at the time of treatment) were known, we would have perfect causal knowledge about the effect of X. Assigning the person p_1 which is in situation s_1 (i.e., unit u_1) to x_1 would yield an expected Y-value 85, whereas assigning unit u_1 to x_2 would yield an expected Y-value 91.

If these numbers are *not* known it is still informative to know the causally unbiased expectations of Y given the two values x_1 and x_2 in the population, i.e.:

$$CUE(Y \mid X = x_1) = \sum_u E(Y \mid X = x_1, U = u)\ P(U = u)$$

$$= 85 \cdot .10 + 74 \cdot .40 + 106 \cdot .10 + 95 \cdot .40 = 86.7$$

and

$$CUE(Y \mid X = x_2) = \sum_u E(Y \mid X = x_2, U = u)\ P(U = u)$$

$$= 91 \cdot .10 + 78 \cdot .40 + 112 \cdot .10 + 99 \cdot .40 = 91.1,$$

as well as the difference between these two expected values, the average causal effect,

$$ACE(1,2) = CUE(Y \mid X = x_1) - CUE(Y \mid X = x_2) = 86.7 - 91.1 = -4.4.$$

For a randomly drawn observational unit (a person-in-a-situation), the expected effect of doing x_1 instead of x_2 is −4.4. This is the best guess for the individual causal effect if no information whatsoever about the person and the situation is available.

The Technical Parameters. In this example, too, we can illustrate that the parameters computed above can be obtained without bias only in very special cases (see Theorems 1 and 2, for instance). In the case of unequal individual assignment probabilities (Example IIa) the conditional expected values are *not* causally unbiased:[23]

$$E(Y \mid X = x_1) = \textstyle\sum_u E(Y \mid X = x_1,\ U = u)\ P(U = u \mid X = x_1)$$

$$= 85 \cdot .043 + 74 \cdot .085 + 106 \cdot .191 + 95 \cdot .681 \approx 94.9,$$

$$E(Y \mid X = x_2) = \textstyle\sum_u E(Y \mid X = x_2,\ U = u)\ P(U = u \mid X = x_2)$$

$$= 91 \cdot .151 + 78 \cdot .679 + 112 \cdot .019 + 99 \cdot .151 \approx 83.8.$$

Obviously, $E(Y \mid X = x_1) \neq CUE(Y \mid X = x_1)$ and $E(Y \mid X = x_2) \neq CUE(Y \mid X = x_2)$, and

$$PFE(1,2) = 94.9 - 83.8 = 11.1 \neq ACE(1,2).$$

In fact, in Example IIa, the prima facie effect is positive, namely 11.1, whereas the average causal effect is negative, namely −4.4.

However, assuming equal assignment probabilities (Examples IIb and IIc), i.e., stochastic independence of U and X, yields causally unbiased conditional expected values. For Examples IIb and IIc we get:

$$E(Y \mid X = x_1) = 85 \cdot .10 + 74 \cdot .40 + 106 \cdot .10 + 95 \cdot .40 = 86.7$$

and

[23] Again, note that the conditional probabilities $P(U = u \mid X = x_1) = P(X = x_1 \mid U = u) \cdot P(U = u) / P(X = x_1)$ have to be computed from the probabilities displayed in Table 5.

$$E(Y \mid X = x_2) = 91 \cdot .10 + 78 \cdot .40 + 112 \cdot .10 + 99 \cdot .40 = 91.1.$$

In these cases $E(Y \mid X = x_1) = CUE(Y \mid X = x_1)$ and $E(Y \mid X = x_2) = CUE(Y \mid X = x_2)$. Hence, with the individual assignment probabilities of Examples IIb and IIc, the prima facie effect $PFE(1,2) = E(Y \mid X = x_1) - E(Y \mid X = x_2) = 86.7 - 91.1 = -4.4$ *is* causally unbiased, i.e., the prima facie effect is equal to the average causal effect $ACE(1,2)$.

To summarize: This example illustrates that there are applications of the theory of individual and average causal effects in which the distribution of the units is not necessarily such that $P(U = u) = 1/N$. Usually, the distribution of the units is unknown if the unit is a person-in-a-situation. Nevertheless, according to Theorem 1, we still know that the conditional expected values $E(Y \mid X = x)$ are causally unbiased if U and X are independent, such as in the randomized experiment in which each unit (i.e., each person-in-a-situation) is assigned to the treatment condition with the same probability for all units. This can easily be achieved even if we do not have any information about the situation in which the person is at the time of treatment assignment. We even do not have to know how many possible situations there are nor how probable they are. We just have to secure that each person-situation combination has the same chance to be assigned to the treatment condition.

6. Example III

The first two examples illustrate the merits of the theory of individual and average causal effects. In Examples Ib, Ic, IIb, and IIc causal unbiasedness was induced by independence of the observational-unit variable U and the treatment variable X. Our third example will show that we may have causal unbiasedness in the total population [i.e., $PFE(i, j) = ACE(i, j)$] although there is neither unit-treatment homogeneity nor independence of U and X. It exemplifies that two confounders may affect the response variable in such a way that the biases cancel each other resulting in unbiased expected values in the total population. In this sense unbiasedness may be incidental. This example will also show that causal unbiasedness in the total population does not imply causal unbiasedness in the subpopulations.

Table 6. *An example in which the prima facie effect in the total population is positive and equal to each individual treatment effect, but in which the prima facie effects in both sex groups are negative*

Persons	Gender	$P(U=u)$	$E(Y\|X=x_1, U=u)$	$E(Y\|X=x_2, U=u)$	$P(X=x_1\|U=u)$
u_1	m	1/4	100	95	5/8
u_2	m	1/4	70	65	7/8
u_3	f	1/4	85	80	1/8
u_4	f	1/4	55	50	3/8

Note: The unconditional probability for treatment assignment is $P(X=x_1) = \sum_{i=1}^{4} P(X=x_1 \mid U=u_i) \cdot P(U=u_i) = 1/2$.

Table 6 displays an example in which the variable W (gender) with values m (e.g., male) and f (female) defines two subpopulations. In this example the prima facie effect $PFE(1,2) = E(Y \mid X=x_1) - E(Y \mid X=x_2) = ACE(1,2)$ is *positive* (+5). However, the population $\Omega_U := \{u_1, u_2, u_3, u_4\}$ may be partitioned into subpopulations *in each of which* the corresponding differences, $E(Y \mid X=x_1, W=m) - E(Y \mid X=x_2, W=m)$ and $E(Y \mid X=x_1, W=f) - E(Y \mid X=x_2, W=f)$, are *negative* (–5).

The conditional expected values $E(Y \mid X=x)$ may be computed by Equation (9):[24]

$$E(Y \mid X=x_1) = 100 \cdot 5/16 + 70 \cdot 7/16 + 85 \cdot 1/16 + 55 \cdot 3/16 = 77.5,$$

$$E(Y \mid X=x_2) = 95 \cdot 3/16 + 65 \cdot 1/16 + 80 \cdot 7/16 + 50 \cdot 5/16 = 72.5.$$

Hence, their difference is $77.5 - 72.5 = 5$. The conditional expected values $E(Y \mid X=x, W=m)$ are:

$$E(Y \mid X=x_1, W=m) = 100 \cdot 5/12 + 70 \cdot 7/12 = 82.5,$$

$$E(Y \mid X=x_2, W=m) = 95 \cdot 9/12 + 65 \cdot 3/12 = 87.5,$$

$$E(Y \mid X=x_1, W=f) = 85 \cdot 3/12 + 55 \cdot 9/12 = 62.5,$$

[24] Again, note that the conditional probabilities $P(U=u \mid X=x_1) = P(X=x_1 \mid U=u) \cdot P(U=u) / P(X=x_1)$ have to be computed probabilities displayed in Table 6.

$$E(Y \mid X = x_2,\ W = f\,) = 80 \cdot 7/12 + 50 \cdot 5/12 = 67.5$$

(see Appendix B for computational details.) Hence, in the subpopulations the corresponding differences (the prima facie effects in the subpopulations) are both −5. These differences are neither equal to the prima facie effect in the total population nor to the average causal effects in the subpopulations. Hence, this is an example in which the prima facie effects in the subpopulations are biased although the prima facie effect in the total population is unbiased.

To summarize: This example shows that we may have causal unbiasedness in the total population [i.e., $PFE(1,2) = ACE(1,2)$], although there is neither unit-treatment homogeneity nor independence of U and X. Furthermore, it shows that causal unbiasedness of the prima facie effects in the total population does not imply causal unbiasedness of the prima facie effects $E(Y \mid X = x_1,\ W = w) - E(Y \mid X = x_2,\ W = w)$ in the subpopulations. Hence, causal unbiasedness may be incidental, i.e., it may be a fortunate coincidence instead of being a consequence of careful experimental design or of a stable empirical phenomenon such as unit-treatment homogeneity.

7. Summary and Discussion

In this paper, we reformulated the theory of individual and average causal effects in terms of classical probability theory. We described the kind of random experiments, i.e., the empirical phenomenon to which the theory refers, defined the concepts of *individual* and *average causal effects*, and studied the relation between these notions and the parameters that may be estimated in samples: the conditional expected values $E(Y \mid X = x_i)$ of the response variable Y in treatment condition x_i. For simplicity, we restricted our discussion to the case where there is no concomitant variable or covariate. We defined the differences $E(Y \mid X = x_i) - E(Y \mid X = x_j)$ between these conditional expected values – the prima facie effects $PFE(i, j)$ – to be *causally unbiased* if the prima facie effect is equal to the average causal effect $ACE(i, j)$. This equation, $PFE(i, j) = ACE(i, j)$, holds necessarily if the observational units are randomly assigned to the two experimental conditions. Thus, the theory justifies and gives us a deeper understanding of the randomized experiment. However, the equation $PFE(i, j) = ACE(i, j)$ also holds in the case of unit-treatment homogeneity, i.e., in the case in which, within each treatment condition, each observational unit has the same expected value.

The theory of individual and average causal effects has been illustrated by three examples. Example I emphasized the crucial role of random assignment of units to treatment conditions. It showed that the prima facie effect $PFE(1,2) = E(Y \mid X = x_1) - E(Y \mid X = x_2)$ can be seriously misleading: *Although the causal effect was negative for each and every individual, the prima facie effect was positive.* Other examples can easily be constructed in which the prima facie effect is zero although each individual causal effect is positive (negative).[25] This proves that many textbooks are wrong maintaining that "no correlation is a proof for no causation" or that "correlation is necessary but not sufficient for causation" (see, e.g., Bortz, 1999, p. 226). Example I also illustrates the distinction between the conditional expected value $E(Y \mid X = x)$ and the causally unbiased expected value $CUE(Y \mid X = x)$ of Y given x. This distinction is of utmost importance to empirical science, because it implies that the prima facie effect $PFE(i, j) = E(Y \mid X = x_i) - E(Y \mid X = x_j)$, which is the focus of statistical analyses, may be completely misleading if it does not coincide with the average causal effect $ACE(i, j) = CUE(Y \mid X = x_i) - CUE(Y \mid X = x_j)$. If $E(Y \mid X = x_i) \neq CUE(Y \mid X = x_i)$, the conditional expected value $E(Y \mid X = x_i)$ will be causally biased. This implies that the sample means systematically estimate the wrong parameters. This situation may occur if the individual treatment assignment probabilities depend on the person and its attributes. There can be no bias, however, if the individual treatment assignment probabilities are the same for all units (randomization) or in the case of unit-treatment homogeneity.

Example II showed that there are applications in which the observational units are not persons but persons-in-a-situation. In this case *the observational units do not have a uniform distribution* with $P(U = u) = 1/N$ for each unit. The original formulation of the theory by Rubin and others were restricted to this kind of distribution. Furthermore, in this example, the distribution of the observational units (the persons-in-a-situation) is usually unknown in empirical applications. Nevertheless, according to Theorem 1, we still know that the conditional expected values $E(Y \mid X = x)$ are unbiased if U and X are independent. Again, this independence is guaranteed in the randomized experiment even if we do not have any information about the person's situation at the time of treatment assignment.

---

[25] Simply replace the individual treatment-assignment probabilities in Table 1 by $P(X = x_1 \mid U = u) = 1/4$ for u_1 and $P(X = x_1 \mid U = u) = 1/2$ for u_2. In this case the prima facie effect (and the correlation between X and Y) will be zero, although each and every individual causal effect is negative.

Example III demonstrates that *causal unbiasedness of prima facie effects may be incidental*. Specifically it is shown that although $PFE(i, j) = ACE(i, j)$ holds in the total population, the corresponding equations may not hold in any subpopulation. Hence, prima facie effects in the subpopulations might be seriously biased although they are causally unbiased in the total population. This result is surprising because it is at odds with our intuition.

Which are the merits and which are the limitations of the theory of individual and average causal effects? A *first merit* are the clear-cut and simple definitions of individual and average causal effects. These definitions show that causality *is* amenable to mathematical formalization and reasoning. Obviously, we can go far beyond the saying that "correlation is not causality". The theory makes clear – and this is a *second merit* – under which circumstances conditional expected values estimated by sample means are known to be of substantive interest in empirical causal research: independence of units and treatment conditions and/or unit-treatment homogeneity.

Which are the limitations? The *first limitation* is that the claim of unbiasedness of the prima facie effect (i.e., of the difference between two conditional expected values) in the total population, though well-defined, is not empirically falsifiable. Postulating that the prima facie effect $PFE(i, j)$ is equal to the average causal effect $ACE(i, j)$ or that the conditional expected values $E(Y|X = x)$ are equal to the unbiased expected value $CUE(Y|X = x)$ of Y given x does not imply anything one could show to be wrong in an empirical application. The equation $E(Y|X = x) = CUE(Y|X = x)$ cannot be tested, because $CUE(Y|X = x)$ cannot be estimated unless (a) each individual u has the same probability to be assigned to $X = x$ (i.e., U and X are stochastically independent such as in the randomized experiment) or unless (b) $E(Y|X, U) = E(Y|X)$ (i.e., if there is unit-treatment homogeneity).

Whereas randomization *guarantees* independence of U and X (see Footnote 17), things are much more complicated in nonrandomized studies: We may either directly assume $E(Y|X = x) = CUE(Y|X = x)$ and $PFE(i, j) = ACE(i, j)$ (unbiasedness) without being able to empirically test this assumption; or, we may assume that the sufficient condition "U and X independent and/or unit-treatment homogeneity" holds and empirically test this assumption. However, since this condition is sufficient but not necessary, rejection of this assumption does not disprove unbiasedness.

The *second limitation* is that causal unbiasedness of the prima facie effect in the total population does not imply causal unbiasedness of the prima facie effects in any subpopulation. Hence, the prima facie effects in the subpopulations may be seriously biased although they are unbiased in the total population. This means, for instance, in a 2 × 2 factorial design (e.g., treatment by gender) that we might causally interpret the (prima facie) treatment effect $E(Y \mid X = x_1) - E(Y \mid X = x_2)$ in the total population, but we would go wrong if we would causally interpret the corresponding prima facie effects $E(Y \mid X = x_1, W = w) - E(Y \mid X = x_2, W = w)$ within the gender subpopulations of males ($W = m$) or females ($W = f$).

A *third limitation* is that the concept of causality associated with causally unbiased expected values and average causal effects is rather weak. As mentioned before, even if the average causal effect [see Eq. (3)] is positive, there can be observational units and subpopulations for which the individual or average causal effects are negative. Investigating questions like these is the focus of analyzing interactions in the analysis of variance or of moderator models in regression analyses. A stronger concept would require invariance of the individual effects (see, e.g., Steyer, 1985, 1992, Ch. 9) or at least invariance of the individual effects in subpopulations (Steyer, 1992, Ch. 14). However, although the search for invariant individual causal effects within subpopulations is certainly a fruitful goal, it should be noted that there is no sufficient condition for such an invariance of individual causal effects that could deliberately be created by the experimenter. In contrast, there *is* a sufficient condition for $PFE(i, j) = ACE(i, j)$ that *is* under the control of the experimenter, namely independence of U and X via random assignment of units to treatment conditions. Nevertheless, the search for interaction or moderator effects is important although it does not replace randomization. Instead, it may supplement it in an important way: Randomization guarantees that the prima facie effects in the subpopulations are at least average causal effects, and if the subpopulations were in fact homogeneous, the average effects in the subpopulations were also effects for all the individuals in that subpopulation.

To summarize, we have learned what we are looking for as substantive scientists: individual causal effects or at least average causal effects. Average causal effects are of interest in the total population but also in subpopulations which provide more detailed information on the effects of a treatment variable. We have also learned that average causal effects can unbiasedly be estimated under special circumstances: random assignment of units to treatment conditions and/or unit-treatment homogeneity.

What to do if random assignment of units to treatment conditions is not possible in a specific application? Give up our substantive interest in estimating the average causal effects? Hoping that the conditional expected values $E(Y \mid X = x)$ and their differences are unbiased but not being able to falsify this hypothesis? Testing the sufficient (but not necessary) conditions for causal unbiasedness?

Obviously, the theory of individual and average causal effects provides a number of useful concepts and is able to answer many questions related to randomized experiments. However, questions concerning causal modeling in *nonrandomized* experiments can not be settled altogether by this theory unless it is complemented in an appropriate way.

References

[1] Bauer, H. (1981). *Probability theory and elements of measure theory.* New York: Academic Press.

[2] Cook, T. D. & Campbell, D. T. (1979*). Quasi-experimentation: design and analysis issues for field settings*. Boston: Houghton Mifflin.

[3] Dawid, A. P. (1979). Conditional independence in statistical theory*. Journal of the Royal Statistical Society*, Series B, 41, 1-31.

[4] Dawid, A. P. (1997). Causal inference without counterfactuals. Research Report No. 188, Department of Statistical Science, University College London.

[5] Dudley, R. M. (1989). *Real analysis and probability.* Pacific Grove, CA: Wadsworth & Brooks/Cole.

[6] Holland, P. (1986). Statistics and causal inference (with comments). *Journal of the American Statistical Association*, *81*, 945-970.

[7] Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18, 449-484.

[8] Holland, P. W. & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (eds.), *Principals of modern psychological measurement* (pp. 3-25). Hillsdale, NJ: Erlbaum.

[9] Holland, P. W. & Rubin, D. B. (1988). Causal inference in retrospective studies. Evaluation Review, 23, 203-231.

[10] Neyman, J. (with Iwaszkiewicz, K., and Kolodziejczyk, S.). (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement of Journal of the Royal Statistical Society*, 2, 107-180.

[11] Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5, 465-472.

[12] Pearl, J. (1995). Causal diagrams for experimental research. *Biometrika*, 82, 669-710.

[13] Pearl, J. (1998). Why there is no statistical test for confounding, why many think there is, and why they are almost right. (Technical Report R-256, January 1998).

[14] Pearl, J. (2000). *Causality – Models, reasoning, and inference.* Cambrigde: University Press.

[15] Pedhazur, E. J. & Pedhazur Schmelkin, L. (1991). *Measurement, design, and analysis. An integrated approach.* Hillsdale, NJ: Lawrence Earlbaum.

[16] Pratt, J. W. & Schlaifer, R. (1988). On the interpretation and observation of laws. *Journal of Econometrics*, *39*, 23-52.

[17] Rosenbaum, P. R. (1984a). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, *79*, 565-574.

[18] Rosenbaum, P. R. (1984b). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A*, *147*, 656-666.

[19] Rosenbaum, P. R. (1984c). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, *79*, 41-48.

[20] 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*, *45*, 212-218.

[21] Rosenbaum, P. R. & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

[22] 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.

[23] Rosenbaum, P. R. & Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics*, *41*, 103-116.

[24] Rosenbaum, P. R. & Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33-38.

[25] Rubin, D. (1973a). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, *29*, 185-203.

[26] Rubin, D. B. (1973b). Matching to remove bias in observational studies. *Biometrics*, *29*, 159-183.

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

[28] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

[29] Rubin, D. B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1-26.

[30] Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*, 34-58.

[31] Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics*, *2*, 463-472.

[32] Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association*, *81*, 961-962.

[33] Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472-480.

[34] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, *Series B*, *13*, 238-241.

[35] Sobel, M. E. (1994). Causal inference in latent variables analysis. In A. von Eye & C. C. Clogg (eds.), *Latent variables analysis* (pp. 3-35). Thousand Oaks, CA: Sage.

[36] Sobel, M. E. (1995). Causal inference in the Social and Behavioral Sciences. In G. Arminger, C. C. Clogg & M. E. Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 1-38). New York: Plenum.

[37] Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* New York: Springer.

[38] Steyer, R. (1985). Causal regressive dependencies: an introduction. In J. R. Nesselroade & A. von Eye (ed.), *Individual development and social change: explanatory analysis* (pp. 95-124). Orlando, FL: Academic Press.

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

[40] Steyer, R., Ferring, D. & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*, 79-98.

[41] Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, 13, 389 – 408.

[42] Steyer, R., Gabler, S., von Davier, A.& Nachtigall, C. (in press). Causal regression models II: unconfoundedness and causal unbiasedness. *Methods of Psychological Research – online.*

[43] Williams, D. (1991). *Probability with martingales.* Cambridge: Cambridge University Press.

[44] Whittaker, J. (1990). *Graphical models in applied multivariate statistics.* Chichester: Wiley.

[45] Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, *2*, 121-134.

Appendix A: Proofs

Proof of Corollary 1. This corollary follows almost directly from Definition 5: Since the sum of a difference is the difference between the sums, we may apply the equation $E(Y|X=x) = CUE(Y|X=x)$ to both terms $E(Y|X=x_i, U=u)$ and $E(Y|X=x_2, U=u)$, hidden in Equation (3), which yields $PFE(i, j) = ACE(i, j)$.

Proof of Theorem 1. Since we assume $P(X=x, U=u) > 0$ for each pair (x, u) of X and U, Equation (9) is always true. If U and X are stochastically independent, then $P(U=u|X=x) = P(U=u)$. This implies Equation (6).

Proof of Theorem 2. $E(Y|X, U) = E(Y|X)$ implies $E(Y|X=x, U=u) = E(Y|X=x)$ for each pair of values x of X and u of U. Hence,

$$\begin{aligned}\sum_u E(Y|X=x, U=u)\, P(U=u) &= \sum_u E(Y|X=x)\, P(U=u)\\ &= E(Y|X=x) \sum_u P(U=u)\\ &= E(Y|X=x), \quad \text{for each value } x \text{ of } X.\end{aligned}$$

Appendix B: Computational Details of Example III

The formula to compute the conditional expected values $E(Y|X=x,\ W=w)$ is:

$$E(Y|X=x,\ W=w) = \sum_u E(Y|X=x,\ U=u,\ W=w)\ P(U=u|X=x,\ W=w),$$

$$= \sum_u E(Y|X=x,\ U=u)\ P(U=u|X=x,\ W=w),$$

for each value x of X and each value w of W, where the conditional probabilities $P(U=u|X=x,\ W=m)$ are displayed in Table 7. Note that $E(Y|X=x,\ U=u,\ W=w) = E(Y|X=x,\ U=u)$, because W (gender) is a function of U (representing the units) in this example.

Table 7. *Conditional probabilities* $P(U=u|\ X=x,\ W=w)$ for the example displayed in Table 6

Persons	Gender	$P(U=u\|X=x_1, W=m)$	$P(U=u\|X=x_2, W=m)$	$P(U=u\|X=x_1, W=f)$	$P(U=u\|X=x_2, W=f)$
u_1	m	5/12	9/12	0	0
u_2	m	7/12	3/12	0	0
u_3	f	0	0	3/12	7/12
u_4	f	0	0	9/12	5/12

Note: Since $W=f(U)$, where f maps Ω_U into the set $\{m, f\}$, the conditional probabilities displayed in this table can be computed from the conditional probabilities $P(U=u|X=x)$. The latter can be computed from the conditional probabilities $P(X=x\ |U=u)$ (see Table 6). The formula for $P(U=u_1|X=x_1,\ W=m)$ is $P(U=u_1|X=x_1,\ W=m) = P(U=u_1|\ X=x_1)\ /\ [P(U=u_1|\ X=x_1) + P(U=u_2|X=x_1)]$.