Chapter for R. H. Hoyle (Ed.), Handbook of Structural Equation Modeling. New York: Guilford Press.

TECHNICAL REPORT
R-370
March 2011

The Causal Foundations of Structural Equation Modeling

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

March 15, 2011

1 Introduction

The role of causality in SEM research is widely perceived to be, on the one hand, of pivotal methodological importance and, on the other hand, confusing, enigmatic and controversial. The confusion is vividly portrayed, for example, in the influential report of Wilkinson and Task Force's (1999) on "Statistical Methods in Psychology Journals: Guidelines and Explanations." In discussing SEM, the report starts with the usual warning: "It is sometimes thought that correlation does not prove causation but 'causal modeling' does. [Wrong! There are] dangers in this practice." But then ends with a startling conclusion: "The use of complicated causal-modeling software [read SEM] rarely yields any results that have any interpretation as causal effects." The implication being that the entire enterprise of causal modeling, from Sewell Wright (1921) to Blalock (1964) and Duncan (1975), the entire literature in econometric research, including modern advances in graphical and nonparametric structural models have all been misguided, for they have been chasing parameters that have no causal interpretation.

The motives for such overstatements notwithstanding, readers may rightly ask: "If SEM methods do not 'prove' causation, how can they yield results that have causal interpretation?" Put another way, if the structural coefficients that SEM researchers labor to estimate can legitimately be interpreted as causal effects then, unless these parameters are grossly misestimated, why deny SEM researchers the honor of "establishing causation" or at least of deriving some useful claims about causation.

The answer is that a huge logical gap exists between "establishing causation," which requires careful manipulative experiments, and "interpreting parameters as causal effects," which may be based on firm scientific knowledge or on previously conducted experiments, perhaps by other researchers. One can legitimately be in a possession of a parameter that stands for a causal effect and still be unable, using statistical means alone, to determine the magnitude of that parameter given non-experimental data. As a matter of fact, we know that

no such statistical means exists; that is, causal effects in observational studies can only be substantiated from a combination of data and untested, theoretical assumptions; not from the data alone. Thus, if reliance on theoretical assumptions disqualifies SEM's parameters from having an interpretation as causal effects, no method whatsoever can endow any parameter with such interpretation, and causal vocabulary should be purged from scientific discourse – an unthinkable restriction.

But then, if the parameters estimated by SEM methods are legitimate carriers of causal claims, and if those claims cannot be proven valid by the data alone, what is the empirical content of those claims? What good are the numerical values of the parameters? Can they inform prediction, decision or scientific understanding? Are they not merely fiction of one's fancy, comparable say to horoscopic speculations?

The aim of this chapter is to lay a coherent logical framework for answering these foundational questions. Following a brief historical account of how the causal interpretation of SEM was obscured (Section 2), we will explicate the empirical content of SEM's claims (Section 3), and describe the tools needed for solving most (if not all) problems involving causal relationships (Sections 4 and 5). The tools are based on nonparametric structural equation models – a natural generalization of those used by econometricians and social scientists in the 1950-60s, which will serve as an Archimedean Point to liberate SEM from its parametric blinders and illucidate its causal content.

In particular we will introduce

1. Tools of reading and explicating the causal assumptions embodied in SEM models as well as the set of assumptions that support each individual causal claim.

2. Methods of identifying the testable implications (if any) of the assumptions in (1), and ways of testing, not the model in its entirety, but the testable implications of the assumptions behind each individual causal claim.

3. Methods of deciding, prior to taking any data, what measurements ought to be taken, whether one set of measurements is as good as another, and which measurements tend to bias our estimates of the target quantities.

4. Methods for devising critical statistical tests by which two competing theories can be distinguished.

5. Methods of deciding mathematically if the causal relationships of interest are estimable from the data and, if not, what additional assumptions, measurements or experiments would render them estimable,

6. Methods of recognizing and generating equivalent models which solidify, extend, and amend the heuristic methods of Stelzl (1986) and Lee and Hershberger (1990)

7. Generalization of SEM to categorical data and non-linear interactions, including a solution to the so called "Mediation Problem," (Baron and Kenny, 1986; MacKinnon, 2008).

2 SEM and Causality: A Brief History of Unhappy Encounters

The founding fathers of SEM, from Sewall Wright (1923) and the early econometricians (Haavelmo, 1943; Simon, 1953; Marschak, 1950; Koopmans, 1953), to Blalock (1964) and Duncan (1975) have all considered SEM a mathematical tool for drawing causal conclusions from a combination of observational data and theoretical assumptions. They were explicit about the importance of the latter, but also adamant about the unambiguous causal reading of the model parameters, once the assumptions are substantiated.

In time, however, the causal reading of structural equation models and the theoretical basis on which it rests became suspect of ad hockery, even to seasoned workers in the field. This occurred partially due to the revolution in computer power, which made workers "lose control of their ability to see the relationship between theory and evidence" (Sørensen, 1998, p. 241), and partly due to a steady erosion of the basic understanding of SEMs which Pearl (2009, p. 138) attributes to notational shortsightedness; i.e., the failure of the equality sign to distinguish structural from regressional equations.

In his critical paper of SEM, Freedman (1987, p. 114) challenged the causal interpretation of SEM as "self-contradictory," and none of the 11 discussants of his paper were able to detect his error and to articulate the correct, noncontradictory interpretation of the example presented by Freedman. Instead, SEM researchers appeared willing to accept contradiction as a fundamental flaw in causal thinking, which must always give way to statistical correctness. In his highly cited commentary on SEM, Chin (1998) surrenders to the critics: "researchers interested in suggesting causality in their SEM models should consult the critical writing of Cliff (1983), Freedman (1987), and Baumrind (1993)."

This, together with the steady influx of statisticians into the field, has left SEM researchers in a quandary about the meaning of the SEM parameters, and has caused some to avoid causal vocabulary altogether and to regard SEM as an encoding of parametric family of density functions, void of causal interpretation. Muthén (1987), for example, wrote "It would be very healthy if more researchers abandoned thinking of and using terms such as cause and effect" (Muthén, 1987). Many SEM textbooks have subsequently considered the word "causal modeling" to be an outdated misnomer (e.g., Kelloway, 1998, p. 8), giving clear preference to causality-free nomenclature such as "covariance structure," "regression analysis," or "simultaneous equations." A popular 21st century textbook reaffirms: "Another term that readers may have encountered is causal modeling, which is used mainly in association with the techniques of path analysis. This expression may be somewhat dated, however, as it seems to appear less often in the literature nowadays" (Kline, 2005, p. 9).

Relentless assaults from the potential-outcome paradigm (Rubin, 1974) have further eroded confidence in SEM's adequacy to serve as a language for causation. Sobel (1996), for example, states that the interpretation of the parameters of SEM model as effects "do not generally hold, even if the model is correctly specified and a causal theory is given." Comparing structural equation models to the potential-outcome framework, Sobel (2008) asserts that "In general (even in randomized studies), the structural and causal parameters are not equal, implying that the structural parameters should not be interpreted as effect." Remarkably, formal analysis proves the exact opposite: structural and causal parameters are

one and the same thing, and they should *always* be interpreted as effects (Galles and Pearl, 1998; see Section 4).

Paul Holland, another advocate of the potential-outcome framework, unravels the root of the confusion: "I am speaking, of course, about the equation: $\{y = a + bx + \epsilon\}$. What does it mean? The only meaning I have ever determined for such an equation is that it is a shorthand way of describing the conditional distribution of $\{y\}$ given $\{x\}$" (Holland, 1995, p. 54). We will see that the structural interpretation of the equation above has in fact nothing to do with the conditional distribution of $\{y\}$ given $\{x\}$; rather, it conveys counterfactual information that is orthogonal to the statistical properties of $\{x\}$ and $\{y\}$ (Section 4.4).

We will further see (Section 4.5) that the SEM language in its nonparametric form offers a mathematically equivalent alternative to the potential-outcome framework that Holland and Sobel advocate for causal inference – a theorem in one is a theorem in another. SEM provides in fact the formal mathematical basis from which the potential-outcome notation draws its legitimacy. This, together with its friendly conceptual appeal and effective mathematical machinery explains why SEM retains its status as the prime language for causal and counterfactual analysis.[1] These capabilities are rarely emphasized in standard SEM texts, where they have been kept dormant in the thick labyrinths of software packages, goodness-of-fit measures, linear regression, MLE estimates, and other details of parametric modeling. The nonparametric perspective unveils these potentials and avails them for both linear and nonlinear analyses.

3 The Logic of SEM

Trimmed and compromised by decades of statistical assaults, textbook descriptions of the aims and claims of SEM grossly understate the power of the methodology. Byrne (2006) for example, describes SEM as "as statistical methodology that takes a confirmatory (i.e., hypothesis-testing) approach to the analysis of a structural theory bearing on some phenomenon. . . The hypothesized model can then be tested statistically in a simultaneous analysis of the entire system of variables to determine the extent to which it is consistent with the data. If goodness-of-fit is adequate, the model argues for the plausibility of postulated relations among variables; if it is inadequate, the tenability of such relations is rejected."

Taken literally, this confirmatory approach encounters some basic logical difficulties. Consider, for example, the hypothesized model:

$$M = \text{"Cinderella is a terrorist"}$$

Although, goodness-of-fit tests with any data would fail to uncover inconsistency in this hypothesized model, we would find it odd to argue for its plausibility. Attempts to repair the argument by insisting that M be falsifiable and invoke only measured variables does not remedy the problem. Choosing

$M =$ "Barameter readings cause rain and the average age in Los Angeles is higher than 3"

[1] A more comprehensive account of the history of SEM and its causal interpretations is given in Pearl (1998). Pearl (2009, pp. 368–74) devotes a section of his book *Causality* to advise SEM students on the causal reading of SEM and how do defend it against the skeptics.

will encounter a similar objection; although M is now falsifiable, and all its variables measured, its success in fitting the data tells us nothing about the causal relations between rain and barometers.

The only way to avoid this paradox is to insist that the tested component of M (that the average age is higher than 3) be logically related to its claims (that barometers cause rain), but this stands contrary to the philosophy of confirmatory analysis, according to which the hypothesized model is submitted to a test "of the entire system of variables," irrespective of whether the tested part bears any relationship to the resulting claims.

This simple, albeit contrived example, uncovers a basic logical flaw in the conservative confirmatory approach, and underscores the need to spell out the empirical content of the assumptions behind the hypothesized model, the claims inferred by the model, and the degree to which data corroborate the latter.

The interpretation of SEM methodology that emerges from the nonparametric perspective (Pearl, 2009, pp. 159–63, 368–74), makes these specifications explicit and is, therefore, free of such flaws. According to this interpretation, SEM is an inference method that takes three inputs and produces three outputs. The inputs are:

I-1. A set A of qualitative causal *assumptions* which the investigator is prepared to defend on scientific grounds, and a model M_A that encodes these assumptions. (Typically, M_A takes the form of a path diagram or a set of structural equations with free parameters. A typical assumption is that certain omitted factors, represented by error terms, are uncorrelated with some variables or among themselves, or that no direct effect exists between a pair of variables.)

I-2. A set Q of *queries* concerning causal and counterfactual relationships among variables of interest. Traditionally, Q concerned the magnitudes of structural coefficient but, in general models, Q will address causal relations more directly, e.g.,

$$Q_1: \quad \text{What is the effect of treatment } X \text{ on outcome } Y?$$

$$Q_2: \quad \text{Is this employer guilty of gender discrimination?}$$

Theoretically, each query $Q_i \in Q$ should be computable from a fully specified model M in which all functional relationships are given. Non-computable queries are inadmissible.

I-3. A set D of experimental or non-experimental *data*, presumably generated by a process consistent with A.

The outputs are

O-1. A set A^* of statements which are the logical implications of A, independent of the data at hand. For example, that X has no effect on Y if we hold Z constant, or that Z is an instrument relative to $\{X, Y\}$.

O-2. A set C of data-based *claims* concerning the magnitudes or likelihoods of the target queries in Q, each conditional of A. C may contain, or example, the estimated mean and variance of a given structural parameter, or the expected effect of a given intervention. Auxiliary to C, SEM also generates an estimand $Q_i(P)$ for each query in Q, or a determination that Q_i is not identifiable from P (Definition 1.)

O-3. A list T of testable statistical implications of A, and the degree $g(T_i), T_i \in T$, to which the data agrees with each of those implications. A typical implication would be the vanishing of a specific partial correlation; such constraints can be read from the model M_A and confirmed quantitatively by the data (Definition 3).

The structure of this inferential exercise is shown schematically in Figure 1.

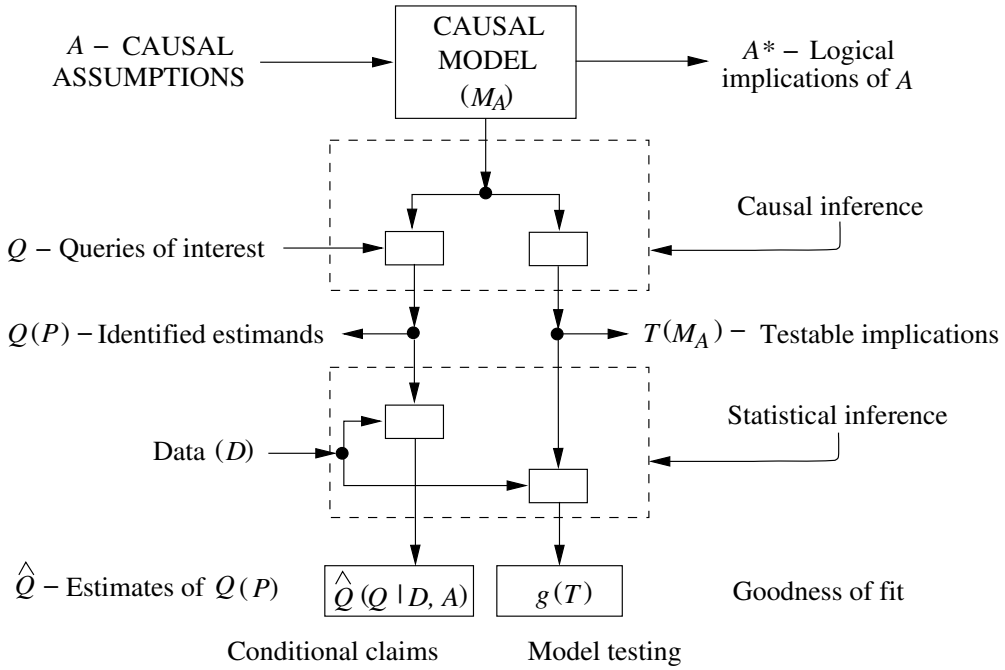

Figure 1: SEM methodology depicted as the an inference engine converting assumptions (A), queries (Q), and data (D) into logical implications (A^*) Conditional claims (C) and data-fitness indices ($g(T)$).

Several observations are worth noting before illustrating these inferences by examples. First, SEM is not a traditional statistical methodology, typified by hypothesis testing or estimation, because neither claims nor assumptions are expressed in terms of probability functions of realizable variables (Pearl, 2009).

Second, all claims produced by an SEM study are conditional on the validity of A, and should be reported in conditional format: "If A then C_i" for any claim $C_i \in C$. Such claims, despite their provisional character, are significantly more assertive than their meek, confirmatory predecessors. They assert that anyone willing to accept A, must also accept C_i out of logical necessity. Moreover, no other method can do better, that is, if SEM analysis finds that a set A of assumptions is necessary for inferring a claim C_i, no other methodology can infer C_i with a weaker set of assumptions.[2]

Thirdly, passing a goodness-of-fit test is not a prerequisite for the validity of the conditional claim "If A then C_i," nor for the validity of C_i. While it is important to know if any

[2]This is important to emphasize in view of often heard critics that, in SEM, one must start with a model in which all causal relations are presumed known, at least qualitatively. Other methods must rest on the same knowledge, though some tend to hide the assumptions under catch-all terms such as "ignorability" or "nonconfoundedness."

assumptions in A are inconsistent with the data, M_A may not have any testable implications whatsoever. In such a case, the assertion "If A then C_i" may still be extremely informative in a decision making context, since each C_i conveys quantitative information extracted from the data rather then qualitative assumptions A with which the study commences. Moreover, even if A turns out inconsistent with D, the inconsistencies may be entirely due to portions of the model which have nothing to do with the derivation of C_i. It is therefore important to identify which statistical implication of (A) is responsible for the inconsistency; global tests for goodness-of-fit hide this information (Pearl, 2009, 2004, pp. 144-45).

Finally, and this has been realized by SEM researchers in the late 1980's, there is nothing in SEM's methodology to protect C from the inevitability of contradictory equivalent models, namely, models that satisfy all the testable implications of M_A and still advertise claims that contradict C. Modern developments in graphical modeling have devised visual and algorithmic tools for detecting, displaying, and enumerating equivalent models. Researchers should keep in mind therefore that only a tiny portion of the assumptions behind each SEM study lends itself to scrutiny by the data; the bulk of it must remain untestable, at the mercy of scientific judgment.

4 The Causal Reading of Structural Equation Models

4.1 The assumptions and their representation

In this section we will illustrate the inferences outlined in Figure 1 using simple structural models consisting of linear equations and their nonparametric counterparts, encoded via diagrams. Consider the linear structural equations

$$y = \beta x + u_Y, \quad x = u_X \tag{1}$$

where x stands for the level (or severity) of a disease, y stands for the level (or severity) of a symptom, and u_Y stands for all factors, other than the disease in question, that could possibly affect Y when X is held constant. In interpreting this equation we should think of a physical process whereby nature *examines* the values of all variables in the domain and, accordingly, *assigns* to variable Y the value $y = \beta x + u_Y$. Similarly, to "explain" the occurrence of disease X, we write $x = u_X$, where U_X stands for all factors affecting X, which may in general include factors in U_Y.

To express the directionality of the underlying process, we should either replace the equality sign with an assignment symbol :=, or augment the equation with a "path diagram," in which arrows are drawn from causes to their effects, as in Figure 2. The absence of an arrow makes the empirical claim that Nature assigns values to one variable irrespective of another. In our example, the diagram encodes the possible existence of (direct) causal influence of X on Y, and the absence of causal influence of Y on X, while the equations encode the quantitative relationships among the variables involved, to be determined from the data. The "path coefficient," β, quantifies the (direct) causal effect of X on Y. Once we commit to a particular numerical value of β, the equation claims that a unit increase for X would result in β units increase of Y regardless of the values taken by other variables in the

model, regardless of the statistics of U_X and U_Y, and regardless of whether the increase in X originates from external manipulations or variations in U_X.

The variables U_X and U_Y are called "exogenous"; they represent observed or unobserved background factors that the modeler decides to keep unexplained—that is, factors that influence but are not influenced by the other variables (called "endogenous") in the model. Unobserved exogenous variables in structural equations, sometimes called "disturbances" or "errors," differ fundamentally from residual terms in regression equations. The latters, usually denoted by letters ϵ_X, ϵ_Y, are artifacts of analysis which, by definition, are uncorrelated with the regressors. The formers are shaped by physical reality (e.g., genetic factors, socioeconomic conditions), not by analysis; they are treated as any other variable, though we often cannot measure their values precisely and must resign ourselves to merely acknowledging their existence and assessing qualitatively how they relate to other variables in the system.

If correlation is presumed possible, it is customary to connect the two variables, U_Y and U_X, by a dashed double arrow, as shown in Figure 2(b). By allowing correlations among omitted factors, we encode in effect the presence of *latent* variables affecting both X and Y, as shown explicitly in Figure 2(c), which is the standard representation in the SEM literature (e.g., Bollen, 1989). If, however, our attention focuses on causal relations among observed rather than latent variables, there is no reason to distinguish between correlated errors and interrelated latent variables; it is only the distinction between correlated and uncorrelated errors (e.g., between Figure 2(a) and (b)) that need to be made.[3] Moreover, when the error terms are uncorrelated, it is often more convenient to eliminate them altogether from the diagram (as in Figure 7, Section 5), with the understanding that every variable, X, is subject to the influence of an independent disturbance U_X.

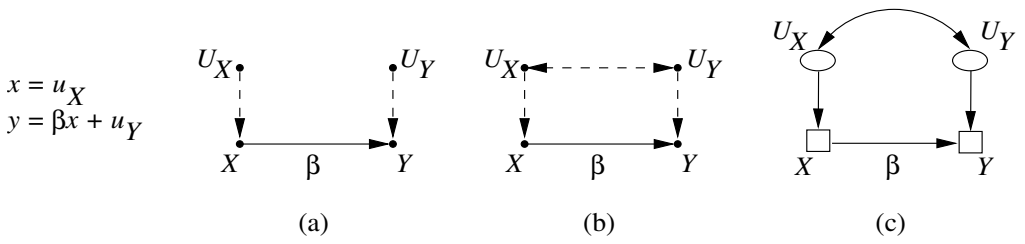

Figure 2: A simple structural equation model, and its associated diagrams, showing (a) independent unobserved exogenous variables (connected by dashed arrows), (b) dependent exogenous variables, and (c) an equivalent, more traditional notation, in which latent variables are enclosed in ovals.

In reading path diagrams, it is common to use kinship relations such as parent, child, ancestor, and descendent, the interpretation of which is usually self-evident. For example, the arrow in $X \rightarrow Y$ designates X as a parent of Y and Y as a child of X. A "path" is any consecutive sequence of edges, solid or dashed. For example, there are two paths between X and Y in Figure 2(b), one consisting of the direct arrow $X \rightarrow Y$ while the other tracing the nodes X, U_X, U_Y, and Y.

[3]Causal relationships among latent variables are assessed by treating their indicators as noisy measurement of the formers (Bollen, 1989; Pearl, 2010c; Cai and Kuroki, 2008).

In path diagrams, causal assumptions are encoded not in the links but, rather, in the missing links. An arrow merely indicates the possibility of causal connection, the strength of which remains to be determined (from data); a missing arrow represents a claim of zero influence, while a missing double arrow represents a claim of zero covariance. Both assumptions are causal, not statistical, since none can be determined from the joint density of the observed variables, X and Y; though both can be tested in experimental settings (e.g., randomized trials).

4.2 Causal Assumptions in Nonparametric Models

To extend the capabilities of SEM methods to models involving discrete variables, nonlinear dependencies, and heterogeneous effect modifications, we need to detach the notion of "effect" from its algebraic representation as a coefficient in an equation, and redefine "effect" as a general capacity to transmit *changes* among variables. The central idea is to exploit the invariant characteristics of structural equations without committing to a specific functional form. For example, the nonparametric interpretation of the diagram in Figure 3(a) corresponds to a set of three unknown functions, each corresponding to one of the observed

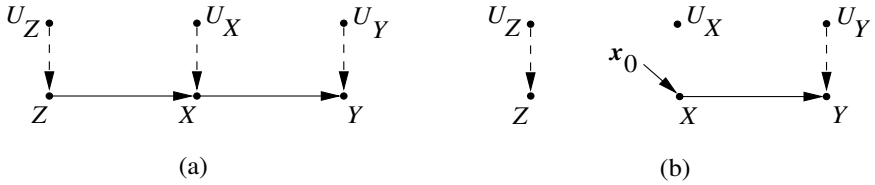

Figure 3: The diagrams associated with (a) the structural model of equation (2) and (b) the modified model of equation (3), representing the intervention $do(X = x_0)$.

variables:

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= f_X(z, u_X) \\ y &= f_Y(x, u_Y), \end{aligned} \tag{2}$$

where in this particular example U_Z, U_X and U_Y are assumed to be jointly independent but otherwise arbitrarily distributed. Each of these functions represents a causal process (or mechanism) that determines the value of the left variable (output) from the values on the right variables (inputs). The absence of a variable from the right-hand side of an equation encodes the assumption that nature ignores that variable in the process of determining the value of the output variable. For example, the absence of variable Z from the arguments of f_Y conveys the empirical claim that variations in Z will leave Y unchanged, as long as variables U_Y and X remain constant.

4.3 Representing Interventions and Causal effects

Remarkably, this feature of invariance permits us to derive powerful claims about causal effects and counterfactuals, despite our ignorance of functional and distributional forms. This

is done through a mathematical operator called $do(x)$, which simulates physical interventions by deleting certain functions from the model, replacing them with a constant $X = x$, while keeping the rest of the model unchanged. For example, to emulate an intervention $do(x_0)$ that holds X constant (at $X = x_0$) in model M of Figure 3(a), we replace the equation for x in equation (2) with $x = x_0$, and obtain a new model, M_{x_0},

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= x_0 \\ y &= f_Y(x, u_Y), \end{aligned} \tag{3}$$

the graphical description of which is shown in Figure 3(b).

The joint distribution associated with the modified model, denoted $P(z, y|do(x_0))$ describes the post-intervention distribution of variables Y and Z (also called "controlled" or "experimental" distribution), to be distinguished from the preintervention distribution, $P(x, y, z)$, associated with the original model of equation (2). For example, if X represents a treatment variable, Y a response variable, and Z some covariate that affects the amount of treatment received, then the distribution $P(z, y|do(x_0))$ gives the proportion of individuals that would attain response level $Y = y$ and covariate level $Z = z$ under the hypothetical situation in which treatment $X = x_0$ is administered uniformly to the population.

In general, we can formally define the postintervention distribution by the equation

$$P_M(y|do(x)) = P_{M_x}(y) \tag{4}$$

In words: In the framework of model M, the postintervention distribution of outcome Y is defined as the probability that model M_x assigns to each outcome level $Y = y$. From this distribution, which is readily computed from any fully specified model M, we are able to assess treatment efficacy by comparing aspects of this distribution at different levels of x_0. However, the central question in the analysis of causal effects is the question of *identification* in partially specified models: Given assumptions set A (as embodied in the model), can the controlled (postintervention) distribution, $P(Y = y|do(x))$, be estimated from data governed by the preintervention distribution $P(z, x, y)$?

In linear parametric settings, the question of identification reduces to asking whether some model parameter, β, has a unique solution in terms of the parameters of P (say the population covariance matrix). In the nonparametric formulation, the notion of "has a unique solution" does not directly apply since quantities such as $Q(M) = P(y|do(x))$ have no parametric signature and are defined procedurally by simulating an intervention in a causal model M, as in equation (3). The following definition captures the requirement that Q be estimable from the data:

Definition 1 (identifiability **)** (Pearl, 2000, p. 77)
A quantity $Q(M)$ is identifiable, given a set of assumptions A, if for any two models M_1 and M_2 that satisfy A, we have

$$P(M_1) = P(M_2) \Rightarrow Q(M_1) = Q(M_2) \tag{5}$$

In words, the functional details of M_1 and M_2 do not matter; what matters is that the assumptions in A (e.g., those encoded in the diagram) would constrain the variability of those details in such a way that equality of P's would entail equality of Q's. When this happens, Q depends on P only and should therefore be expressible in terms of the parameters of P. Section 5.3 will exemplify and operationalize this notion.

4.4 Counterfactual Analysis in Structural Models

Not all questions of causal character can be encoded in $P(y|do(x))$ type expressions, thus implying that not all causal questions can be answered from experimental studies. For example, retrospective questions regarding causes of a given effect (e.g., what fraction of death cases are *due to* a specific treatment) cannot be answered from experimental studies, and naturally this kind of question cannot be expressed in $P(y|do(x))$ notation.[4]

To answer such questions, a probabilistic analysis of counterfactuals is required, one dedicated to the relation "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$. Remarkably, unknown to most economists and philosophers, structural equation models provide the formal interpretation and symbolic machinery for analyzing such counterfactual relationships.[5]

The key idea is to interpret the phrase "had X been x" as an instruction to make a minimal modification in the current model, which may have assigned X a different value, say $X = x'$, so as to ensure the specified condition $X = x$. Such a minimal modification amounts to replacing the equation for X by a constant x, as we have done in equation (3). This replacement permits the constant x to differ from the actual value of X (namely $f_X(z, u_X)$) without rendering the system of equations inconsistent, thus yielding a formal interpretation of counterfactuals in multistage models, where the dependent variable in one equation may be an independent variable in another.

Definition 2 (unit-level counterfactuals) (Pearl, 2000, p. 98)
Let M be a fully specified structural model and M_x a modified version of M, with the equation(s) of X replaced by $X = x$. Denote the solution for Y in the equations of M_x by the symbol $Y_{M_x}(u)$. The counterfactual $Y_x(u)$ (Read: "The value of Y in unit u, had X been x") is given by

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{6}$$

[4]The reason for this fundamental limitation is that no death case can be tested twice, with and without treatment. For example, if we measure equal proportions of deaths in the treatment and control groups, we cannot tell how many death cases are actually attributable to the treatment itself; it is quite possible that many of those who died under treatment would be alive if untreated and, simultaneously, many of those who survived with treatment would have died if not treated.

[5]Connections between structural equations and a restricted class of counterfactuals were first recognized by Simon and Rescher (1966). These were later generalized by Balke and Pearl (1995), using surgeries (equation 6), thus permitting endogenous variables to serve as counterfactual antecedents. The "surgery definition" was used in Pearl (2000, p. 417) and criticized by Cartwright (2007) and Heckman (2005); see Pearl (2009, pp. 362–63, 374–79) for rebuttals.

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x.

We see that every structural equation, say $y = a + bx + u_Y$, carries counterfactual information, $Y_{xz}(u) = q + bx + u_Y$, where Z is any set of variables that do not appear on the right hand side of the equation. Naturally, when U is a random variable, Y_x will be a random variable as well, the distribution of which is dictated by both $P(u)$ and the model M_x. It can be shown (Pearl, 2009, Ch. 7) that equation (6) permits us to define joint distributions of counterfactual variables and to detect conditional independencies of counterfactuals directly from the path diagram.

4.4.1 Reading counterfactuals – An example

This capacity of structural equations to encode and deliver counterfactual information, at both the unit and population levels, is hardly known among SEM researchers, and should receive much greater emphasis in education and the mainstream literature. It is an essential tool to ward off critiques who view counterfactuals as an exclusive property of the potential-outcome framework (Holland, 1988; Wilkinson et al., 1999; Rubin, 2004; Sobel, 2008; Imbens, 2010). This capacity can be demonstrated by a simple example, using a 3-variable linear model; the same one used by Holland (1988) and Sobel (2008) to "prove" that structural models do not have causal or counterfactual content.

Consider the model in Figure 4 where X stands for the level of assistance (or "treatment") given to a student, Z stands for the amount of time the student spends studying, and Y, the outcome, stands for the student's performance on an exam. Starting at a unit level analysis,

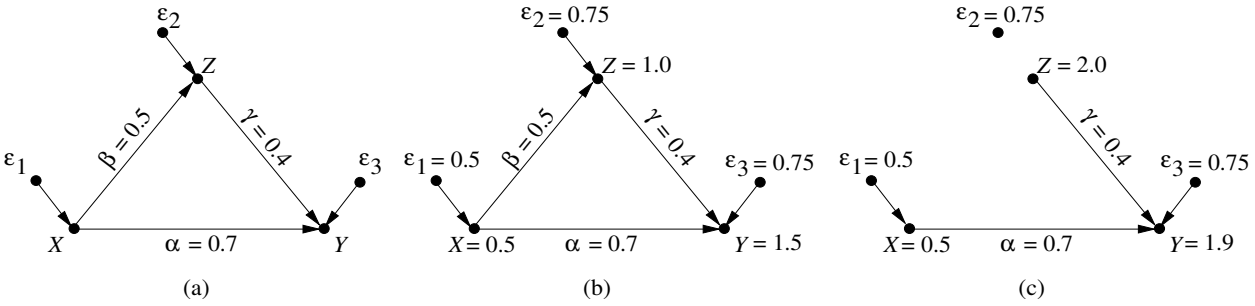

Figure 4: Structural models used for answering a counterfactual question about an individual $u = (\epsilon_1, \epsilon_2, \epsilon_3)$. (a) the generic model, (b) the u-specific model. (c) the modified model necessary to accommodate the antecedent $Z = 2$ of the counterfactual question Q_1.

let us consider a student named Joe, for whom we measure $X = 0.5, Z = 1, Y = 1.5$, and about whom we ask a counterfactual question:

Q_1**:** What would Joe's score be had he doubled his study time?

Using our subscript notation, this question amounts to evaluating $Y_{Z=2}(u)$, with u standing for the distinctive characteristics of Joe, namely, $u = (\epsilon_1, \epsilon_2, \epsilon_3)$, as inferred from the observed data $\{X = 0.5, Z = 1, Y = 1.5\}$.

The answer to this question is obtained in three steps.

1. Use the data to compute the exogenous factors $\epsilon_1, \epsilon_2, \epsilon_3$ (These are the invariant characteristics of unit u, and do not change by interventions or counterfactual hypothesizing.) In our model, we get (Figure 4(b)):

$$\begin{aligned}\epsilon_1 &= 0.5\\ \epsilon_2 &= 1 - 0.5 \times 0.5 = 0.75,\\ \epsilon_3 &= 1.5 - 0.5 \times 0.7 - 1 \times 0.4 = 0.75\end{aligned}$$

2. Modify the model, to form $M_{Z=2}$, in which Z is set to 2 and all arrows to Z are removed (Figure 4(c)).

3. Compute the value of Y in the mutilated model formed in step 2, giving:

$$Y_{Z=2} = 0.5 \times 0.7 + 2.0 \times 0.4 + 0.75 = 1.90$$

This example illustrates the need to modify the original model (Figure 4(a)), in which the combination $(X = 1, \epsilon_2 = 0.75, Z = 2.0)$ constitutes a contradiction (see footnote 5). This is precisely the contradiction that Freedman (1987) could not reconcile in his critic of SEM.

Let us now ask another hypothetical question about Joe.

Q_2**:** What would Joe's score be, had the treatment been 0 and had he studied at whatever level he would have studied had the treatment been 1?

This rather intricate question, which involves nested conditionals, is the basis for defining mediation, to be discussed fully in Section 5.4. Using our subscript notation, the quantity sought can be written as Y_{0,Z_1}, where Z_1 is the value that Z would attain had X been one.

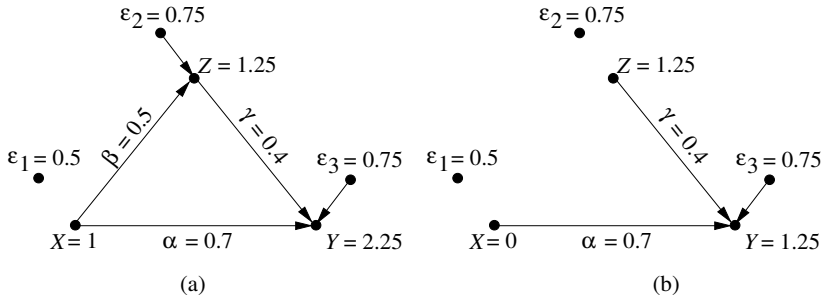

Figure 5: Unit-specific structural models used for answering a nested counterfactual question concerning the indirect effect of X on Y. (a) Modified model needed for calculating Z_1. (b) Modified model needed for calculating Y_{0,Z_1}.

To compute this quantity we need to form two modified models. The first, shown in Figure 5(a), to compute Z_1 and thus provide an answer to Q_2:

$$Z_1 = 1.0 \times 0.5 + 0.75 = 1.25$$

The second, shown in Figure 5(b), to compute Y_{0,Z_1} and thus provide an answer to Q_2:

$$Y_{0,Z_1} = Y_{0,1.25} = 1.25 \times 0.4 + 0.75 = 1.25$$

If we compare this value of $Y_{0,Z_1} = 1.25$ with Joe's outcome had he not received any treatment, $Y_0 = 0.75 \times 0.4 + 0.75, = 1.05$, the difference is, as expected, the indirect effect of X on Y, $Y_{0,Z_1} - Y_0 = 0.20 = \beta \times \gamma$.

This exercise may seem unnecessarily complicated in linear models, where we can compute our desired quantity directly from the product $\beta \times \gamma$. The benefit of using counterfactuals will be revealed in Section 5.5 where indirect effects will be defined for discrete variables, and estimated from data without assuming any parametric forms of the equations.

4.4.2 Predicting outcomes and potential outcomes in empirical studies

Having convinced ourselves that every counterfactual question can be answered (using Eq. (6)) from a fully specified structural model, we next move to population level analysis and ask a policy related question on a set of 10 individuals, each characterized by a distinct vector $u_i = (\epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i})$, as shown in the first three columns of Table 1.

	Participant characteristics			Predicted behavior			Predicted Potential Outcomes				
Participant	ϵ_1	ϵ_2	ϵ_3	X	Y	Z	Y_0	Y_1	Z_0	Z_1	$Y_{00}\ldots$
1	0.5	0.75	0.75	0.5	1.50	1.0	1.05	1.95	0.75	1.25	0.75
2	0.3	0.1	0.4	0.3	0.71	0.25	0.44	1.34	0.1	0.6	0.4
3	0.5	0.9	0.2	0.5	1.01	1.15	0.56	1.46	0.9	1.4	0.2
4	0.6	0.5	0.3	0.6	1.04	0.8	0.50	1.40	0.5	1.0	0.3
5	0.5	0.8	0.9	0.5	1.67	1.05	1.22	2.12	0.8	1.3	0.9
6	0.7	0.9	0.3	0.7	1.29	1.25	0.66	1.56	0.9	1.4	0.3
7	0.2	0.3	0.8	0.2	1.10	0.4	0.92	1.82	0.3	0.8	0.8
8	0.4	0.6	0.2	0.4	0.80	0.8	0.44	1.34	0.6	1.1	0.2
9	0.6	0.4	0.3	0.6	1.04	0.8	0.46	1.36	0.4	0.9	0.3
10	0.3	0.8	0.3	0.3	0.89	0.95	0.62	1.52	0.8	1.3	0.3

Table 1: Potential and Observed Outcomes predicted by the structural model of Figure 4(a).

For each triplet $(\epsilon_1, \epsilon_2, \epsilon_3)$, the model of Figure 4(a) enables us to complete a full row of the table, including Y_0 and Y_1, which stand for the potential outcomes under control ($X = 0$) and treatment ($X = 1$) conditions, respectively. We see that a simple structural model like the one in Figure 4(a) encodes in effect a synthetic population of individuals together with their predicted behavior under both observational and experimental studies. The columns labeled X, Y, Z predict the results of observational studies, and those labeled Y_0, Y_1, Z_0, Z_1 predict the hypothetical outcome under two treatment regimes, $X = 0$, and $X = 1$. Many more, in fact infinite potential outcomes are predicted as well, for example, $Y_{X=0.5,Z=2.0}$ computed in Figure 4(c), and all combinations of subscripted variables. From this synthetic

population one can find the distribution of every counterfactual query on variables X, Y, Z, including, in particular, retrospective counterfactuals, such as the probability that a person chosen at random would have passed the exam by getting assistance given that, in reality, he/she failed the example and did not receive any assistance.[6]

This prediction power was facilitated of course with the help of two untestable pieces of information: (1) the structure of the model (which include the assumption of independent error terms) and (2) the values of the model parameters (which include the distribution of each exogenous variable). Whereas the latter can often be inferred from the data (see Section 5.3), the former depends largely on scientific judgment.

Now assume that we have no information whatsoever about the underlying model and all we have are measurements on Y taken in the experimental study in which X is randomized over two levels, $X = 0$ and $X = 1$.

	Predicted Potential Outcomes		Observed Outcomes	
Participant	Y_0	Y_1	Y_0	Y_1
1	1.05	1.95	1.05	■
2	0.44	1.34	■	1.34
3	0.56	1.46	■	1.46
4	0.50	1.40	■	1.40
5	1.22	2.12	1.22	■
6	0.66	1.56	0.66	■
7	0.92	1.82	■	1.82
8	0.44	1.34	0.44	■
9	0.46	1.36	■	1.36
10	0.62	1.52	0.62	■
	True average treatment effect: 0.90		Study average treatment effect: 0.68	

Table 2: Potential and Observed Outcomes in a randomized clinical trial with X randomized over $X = 0$ and $X = 1$.

Table 2 describes the responses of the same 10 participants (Joe being participant 1) under such experimental conditions. The first two columns give the true potential outcomes (taken from Table 1) while the last two columns describe the information available to the experimenter, where a Square indicates that the response was not observed.[7] Randomization assures us that, although half of the potential outcomes are not observed, the difference

[6]This probability, written $P(Y_1 = 1|X = 0, Y = 0)$, also known as the "probability of causation" (Pearl, 2009, Ch. 9) quantifies "causes of effect," as opposed to "effect of causes," and was excluded, prematurely I presume, from the province of potential outcome analysis (Holland, 1986).

[7]Such tables are normally used to explain the philosophy behind the potential outcome framework (e.g., West and Thoemmes (2010)) in which Y_1 and Y_0 are taken as unexplained random variables. Here they are defined by, and derived from a simple structural model.

between the observed means in the treatment and control groups, will converge to the average of the true difference, $E(Y_1 - Y_0) = 0.9$.

In our model, since all exogenous variables are independent, the slope of the regression of Y on X would also converge to the average causal effect. Bias will be introduced if ϵ_1 is correlated with ϵ_2 or with ϵ_3. However, such correlation will not bias the average causal effect estimated in the experimental study.

4.5 Relations to the Potential Outcome Framework

Definition 2 constitutes the bridge between SEM and a framework called "potential outcome" (Rubin, 1974) which is often presented as a "more principled alternative" to SEM (Holland, 1988; Rubin, 2004; Wilkinson et al., 1999; Sobel, 1996, 2008). Such presentations are misleading and misinformed; the two frameworks have been proven to be a logically equivalent, differing only in the language in which researchers are permitted to express assumptions. A theorem in one is a theorem in the other (Pearl, 2009, pp. 228–31), with Definition 2 providing the formal basis for both.

The idea of potential-outcome analysis is simple. Researchers who feel uncomfortable presenting their assumptions in diagrams or structural equations may do so in a roundabout way, using randomized trial as the ruling paradigm, and interpret the counterfactual $Y_x(u)$ as the potential outcome of subject u to hypothetical treatment $X = x$ ignoring the mechanisms that govern that outcome. The causal inference problem is then set up as one of "missing data," where the missing data are the potential outcomes $Y_x(u)$ under the treatment not received, while the observed data are the potential outcomes under the received treatments, as shown in Table 2.

Thus, Y_x becomes a new latent variable which reveals its value only when $X = x$, through the relation

$$X = x \implies Y_x = Y, \tag{7}$$

sometimes written (for binary X):

$$Y = xY_1 + (1 - x)Y_0$$

Beyond this relation (known as "consistency assumption"), the investigator may ignore the fact that Y_x is actually Y itself, only measured under different conditions (as in Figure 4(c)), and proceed to estimate the average causal effect, $E(Y_{x'}) - E(Y_x)$, with all the machinery that statistics has developed for missing data. Moreover, since (7) is also a theorem in the logic of structural counterfactuals (Pearl, 2009, Ch. 7) and a complete one,[8] researchers in this camp are guaranteed never to obtain results that conflict with those derived in the structural framework.

The weakness of this approach surfaces in the problem formulation phase where, deprived of diagrams and structural equations, researchers are forced to express the (inescapable) assumption set A in a language totally removed from scientific knowledge, for example, in the form of conditional independencies among counterfactual variables (see Pearl, 2010a).

[8] In other words, a complete axiomization of structural counterfactuals in recursive systems consists of (7) and a few non essential details (Halpern, 1998).

For example, to express the fact that, in randomized trial, X is independent on both ϵ_2 and ϵ_3 (Figure 4(a)), the investigator would need to write the cryptic, "strong ignorability" expression $X \perp\!\!\!\perp \{Z_1, Z_0, Y_{00}, Y_{01}, Y_{10}, Y_{11}\}$. To overcome this obstacle, Pearl (2009) has devised a way of combining the best features of the two approaches. It is based on encoding causal assumptions in the language of diagrams or structural equations; translating these assumptions into counterfactual notation; performing derivation in the algebraic language of counterfactuals, using axioms derived from equation (6) and, finally, interpreting the result in plain causal language. The mediation problem discussed in Section 5.4 illustrates how such symbiosis clarifies the conceptualization and estimation of direct and indirect effects, a task that has lingered on for several decades.

5 The Testable Implications of Structural Models

Thus far we discussed the top part of the inference process in Figure 1; i.e., the assumptions that enter into a structural model, the logical implications of those assumptions and typical queries that an investigator may wish to pose, for example, $Q = P(y|do(x))$ or $Q = P(Y_x|x', y', z')$. This section deals with the testable implications of structural models, sometimes called "over-identifying restrictions," and ways of reading them from the graph.

5.1 The *d*-separation criterion

Although each causal assumption in isolation cannot be tested in non-experimental studies, the sum total of all causal assumptions in a model often has testable implications. The chain model of Figure 3(a), for example, encodes seven causal assumptions, each corresponding to a missing arrow or a missing double-arrow between a pair of variables. None of those assumptions is testable in isolation, yet the totality of all seven assumptions implies that Z is unassociated with Y in every stratum of X. Such testable implications can be read off the diagrams using a graphical criterion known as *d-separation* (Pearl, 1988).

Definition 3 (d-separation**)**
A set S of nodes is said to block a path p if either (1) p contains at least one arrow-emitting node that is in S, or (2) p contains at least one collision node that is outside S and has no descendant in S. If S blocks all *paths from set X to set Y, it is said to "d-separate X and Y," and then, X and Y are independent given S, written $X \perp\!\!\!\perp Y|S$.*[9]

To illustrate, the path $U_Z \rightarrow Z \rightarrow X \rightarrow Y$ in Figure 3(a) is blocked by $S = \{Z\}$ and by $S = \{X\}$, since each emits an arrow along that path. Consequently we can infer that the conditional independencies $U_Z \perp\!\!\!\perp Y|Z$ and $U_Z \perp\!\!\!\perp Y|X$ will be satisfied in any probability function that this model can generate, regardless of how we parametrize the arrows. Likewise, the path $U_Z \rightarrow Z \rightarrow X \leftarrow U_X$ is blocked by the null set $\{\emptyset\}$, but it is not blocked by $S = \{Y\}$ since Y is a descendant of the collision node X. Consequently, the marginal independence $U_Z \perp\!\!\!\perp U_X$ will hold in the distribution, but $U_Z \perp\!\!\!\perp U_X|Y$ may or may not hold. This special handling of collision nodes (or *colliders*, e.g., $Z \rightarrow X \leftarrow U_X$) reflects a general

[9] In linear models, d-separation is applicable to cyclic models as well.

phenomenon known as *Berkson's paradox* (Berkson, 1946), whereby observations on a common consequence of two independent causes render those causes dependent. For example, the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail.

The testable implications of any given model are vividly advertised by its associated graph G. Each d-separation condition in G corresponds to a conditional independence test that can be performed on the data to support or refute the validity of M. These can easily be enumerated by attending to each missing edge in the graph. For example, in Figure 6, three

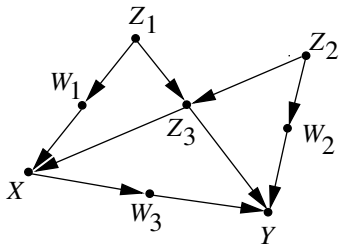

Figure 6: A Markovian model illustrating d-separation. Error terms are assumed mutually independent and not shown explicitly.

of the missing edges are $Z_1 - Z_2, Z_1 - Y$, and $Z_2 - X$. Accordingly, the testable implications of M include:

$$\begin{aligned} Z_1 &\perp\!\!\!\perp Z_2 \\ Z_1 &\perp\!\!\!\perp Y|\{X_1, Z_2, Z_3\} \\ Z_2 &\perp\!\!\!\perp X|\{Z_1, Z_3\}. \end{aligned}$$

In linear systems, these conditional independence constraints translate into zero coefficients in the corresponding regression equations. For example, the three implications above translate into $a = 0, b_1 = 0$, and $c_1 = 0$ in the following regressions:

$$\begin{aligned} Z_1 &= aZ_2 + \epsilon \\ Z_1 &= b_1Y + b_2X + b_3Z_2 + b_4Z_3 + \epsilon' \\ Z_2 &= c_1X + c_3Z_1 + c_4Z_3 + \epsilon''. \end{aligned}$$

Such tests are easily conducted by routine regression techniques, and they provide valuable diagnostic information for model modification, in case any of them fail (see Pearl, 2009, pp. 143–45). Software routines for automatic detection of all such tests, as well as other implications of graphical models, are reported in Kyono (2010).

If the model is Markovian (i.e., acyclic with no unobserved confounders), then the d-separation conditions are the ONLY testable implications of the model. If the model contains latent common causes, then additional constraints are imposed, beyond the d-separation conditions.[10]

[10]These constraints called "dormant independence" (Shpitser and Pearl, 2008) or Verma's constraints (Verma and Pearl, 1990), are given covariant-matrix interpretation in McDonald (2002).

5.2 Equivalent Models

D-separation also defines conditions for model equivalence that are easily ascertained in the Markovian models (Verma and Pearl, 1990) as well a semi-Markovian models (Ali et al., 2009). These mathematically proven conditions should supersede the heuristic (and occasionally faulty) rules prevailing in SEM's research (Lee and Hershberger, 1990).

For example, consider the model of Figure 7. According to the replacement criterion of

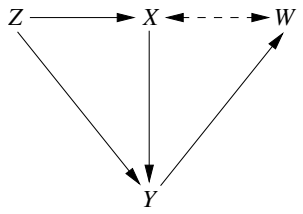

Figure 7: Showing discrepancy between Lee and Hershberger's replacement rule and d-separation, which forbids the replacement of $X \rightarrow Y$ by $X \leftrightarrow Y$.

Lee and Hershberger we can replace the arrow $X \rightarrow Y$ with a double-arrow edge $X \leftrightarrow Y$ (representing residual correlation), since all predictors (Z) of the effect variable (Y) are the same as those for the source variable (X). Unfortunately, the postreplacement model imposes additional constraint, $r_{WZ \cdot Y} = 0$, that is not imposed by the pre-replacement model. This can be seen from the fact that, conditioned on Y, the path $Z \rightarrow Y \leftarrow X \leftrightarrow W$ is unblocked and will becomes blocked if replaced by $Z \rightarrow Y \leftrightarrow X \leftrightarrow W$. The same applies to path $Z \rightarrow X \leftrightarrow W$, since Y would cease to be a descendant of X.

Criteria for recognizing nested models demand similar reform (see chapter on equivalent models in this volume).

5.3 Identification Using Graphs—the Back-Door Criterion

Consider an observational study where we wish to find the effect of X on Y—for example, treatment on response—and assume that the factors deemed relevant to the problem are structured as in Figure 6; some of these factors may be unmeasurable, such as genetic trait or life style; others are measurable, such as gender, age, and salary level. Using the terminology of Section 3, our problem is to determine whether the query $Q = P(y|do(x))$ is identifiable, given the model and, if so, to derive an estimand $Q(P)$ to guide the estimation of Q.

This problem is typically solved by "adjustment," that is, selecting a subset of factors for measurement, so that comparison of treated versus untreated subjects having the same values of the selected factors gives the correct treatment effect in that subpopulation of subjects. Such a set of factors is called a "sufficient set" or "admissible set" for adjustment.

The following criterion, named "back-door" in Pearl (1993), provides a graphical method of selecting admissible sets of factors, and demonstrates that nonparametric queries such as $Q = P(y|do(x))$ can sometimes be identified with no knowledge of the functional form of the equations or the distributions of the latent variables in M.

Definition 4 (admissible sets—the back-door criterion) *A set S is admissible (or "sufficient") if two conditions hold:*

1. *No element of S is a descendant of X.*
2. *The elements of S "block" all "back-door" paths from X to Y—namely, all paths that end with an arrow pointing to X.*

In this criterion, "blocking" is interpreted as in Definition 1. Based on this criterion we see, for example in Figure 6, that the sets $\{Z_1, Z_2, Z_3\}, \{Z_1, Z_3\}$, $\{W_1, Z_3\}$, and $\{W_2, Z_3\}$ are each sufficient for adjustment, because each blocks all back-door paths between X and Y. The set $\{Z_3\}$, however, is not sufficient for adjustment because it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

The intuition behind the back-door criterion is as follows. The back-door paths in the diagram carry spurious associations from X to Y, while the paths directed along the arrows from X to Y carry causative associations. Blocking the former paths (by conditioning on S) ensures that the measured association between X and Y is purely causal, namely, it correctly represents the target quantity: the causal effect of X on Y. The reason for excluding descendants of X (e.g., W_3 or any of its descendants) and conditions for relaxing this restriction are given in (Pearl, 2009, p. 338–41).

5.3.1 Identifying causal effects

The back-door criterion provides a powerful solution to many identification problems, and is summarized in the next theorem.

Theorem 1 *(Causal Effects Identification)*
For any two disjoint sets of variables, X and Y in a causal diagram G, the causal effect of X on Y is given by

$$P(Y = y|do(X = x)) = \sum_s P(Y = y|X = x, S = s)P(S = s) \tag{8}$$

where S is any set of covariates satisfying the back-door condition of Definition 4.

Since all factors on the right-hand side of the equation are estimable (e.g., by regression) from pre-interventional data, the causal effect can likewise be estimated from such data without bias.

In linear systems, identified causal effect expressions like equation (8) reduce to sums and products of partial regression coefficients. For example, if we wish to estimate the total effect τ_{XY} of X on Y in the linear version of Figure 6, we simply take the regression coefficient of Y on X, partialled on any sufficient set S, giving:

$$\tau_{XY} = r_{YX \cdot S} = r_{YX \cdot Z_1, Z_3} = r_{YX \cdot W_1, Z_3} = \ldots$$

Current SEM practices do not take advantage of this capability to decide identification prior to obtaining data, and to estimate the identified quantities directly, by partialling out

sufficient sets. Rather, the prevailing practice is to identify the model in its entirety by running ML routines on noisy data and hoping for their convergence. This is unfortunate (see chapter on identification) because the target quantity may often be identifiable when the model as a whole is not (see (Pearl, 2009, p. 151) for examples). Moreover, estimation accuracy deteriorates when we allow noisy data of irrelevant variables to corrupt the estimation of the target quantity (McDonald, 2004). The theory of d-separation and the back-door criterion enable us to focus the identification of target quantities on the relevant variables and reduce it to algorithmic routines (Kyono, 2010). We also note that, when applied to linear models, all identification conditions are valid for feedback systems as well.

5.3.2 When do structural coefficients equal regression coefficients

Remarkably, a close cousin of the back door criterion, has resolved an age-long identification problem in linear SEMs: Under what conditions can a path coefficient β_{XY} be estimated as a regression coefficient, and what variables should serve as the regressors? The answer is given by a criterion called "single door" (Pearl, 2009, p. 150) which reads:

Corollary 1 *(the single door criterion)*
Let β_{XY} be the structural coefficient labeling the arrow $X \to Y$ and let $r_{YX \cdot S}$ stand for the X coefficient (slope) in the regression of Y on X and S, namely, $r_{YX \cdot S} = \frac{\partial}{\partial x} E(Y|x, s)$. The equality $\beta_{XY} = r_{YX \cdot S}$ holds if

1. *the set S contains no descendant of Y and*
2. *S blocks all paths between X to Y, except the direct path $X \to Y$.*

In Figure 7, for example, β_{XY} equals the coefficient b_1 in the regression $Y = b_1X + b_2Z + \epsilon$. while β_{YW}, labeling the arrow $Y \to W$, is equal to the coefficient c_1 in the regression

$$W = c_1Y + c_2X + c_3Z + \epsilon.$$

Note that regressing W on Y and X alone is insufficient, for it would leave the path $Y \leftarrow Z \to X \leftrightarrow W$ unblocked. In a similar fashion we can obtain

$$\beta_{ZY} = r_{YZ \cdot X}$$
$$\beta_{ZX} = r_{XZ}$$

If no set S can be found that satisfies the conditions of Corollary 1 then β_{XY} cannot be reduced to a single regression coefficient, and other identification techniques may be invoked, for example, instrumental variables.

5.3.3 Recognizing Instrumental Variables

Instrumental variables is one of the oldest identification technique devised for linear systems (Wright, 1928). The method relies on finding a variable Z that is correlated with X and is deemed uncorrelated with the error term in an equation (see Pearl, 2009, pp. 242–48, for formal definition). While no statistical test can certify a variable as instrument, the

d-separation criterion permits us to identify such variables in the causal graph, and use them to identify parameters that do not satisfy the condition of Corollary 1. Moreover, the graph also shows us how to turn variables into instruments when none exist. In Figure 6, for example, Z_1 is not an instrumental variable for the effect of Z_3 on Y, because there is a directed path from Z_3 to Y, via W_1 and X. Controlling for X will not remedy the situation because X being a descendant of Z_3 would unblock the path $Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$. However, controlling for W_1 will render Z_1 a legitimate instrumental variable, since all paths connecting Z_1 to Y would go through Z_3.

The general criterion is given by the following Theorem.

Theorem 2 *(Identification using instrumental variables)*
Let β_{XY} stand for the path coefficient assigned to the arrow $X \rightarrow Y$ in a causal graph G. Parameter β_{XY} is identified if there exists a pair (Z, W), where Z is a single node in G (not excluding $Z = X$), and W is a (possibly empty) set of nodes in G, such that:

1. *W consists of nondescendants of Y,*
2. *W d-separates Z from Y in the graph G_{XY} formed by removing $X \rightarrow Y$ from G,*
3. *Z and X are d-connected, given W, in G_{XY}.*

Moreover, the estimand induced by the pair (Z, W) is given by:

$$\beta_{XY} = \frac{cov(Y, Z|W)}{cov(X, Z|W)}.$$

Additional identification conditions for linear models are given in Pearl (2009, Ch. 5), McDonald (2002, 2004), and Brito and Pearl (2002) and implemented in Kyono (2010). Complete graphical criteria for causal-effect identification in nonparametric models is developed in Tian and Pearl (2002) and Shpitser and Pearl (2006b).

5.4 Mediation: Direct and Indirect Effects

5.4.1 Decomposing effects, aims, and challenges

The decomposition of effects into their direct and indirect components carries theoretical scientific importance, for it tells us "how nature works" and, therefore, enables us to predict behavior under a rich variety of conditions and interventions. For example, an investigator may be interested in assessing the extent to which the effect of a given variable can be reduced by weakening an intermediate process, standing between that variable and the outcome.

Structural equation models provide a natural language for analyzing path-specific effects and, indeed, considerable literature on direct, indirect, and total effects has been authored by SEM researchers (Bollen, 1989)), for both recursive and nonrecursive models. This analysis usually involves sums of powers of coefficient matrices, where each matrix represents the path coefficients associated with the structural equations.

Yet despite its ubiquity, the analysis of mediation has long been a thorny issue in the social and behavioral sciences (Baron and Kenny, 1986; MacKinnon, 2008) primarily because

the distinction between causal parameters and their regressional interpretations were often conflated, as in (Holland, 1995) and (Sobel, 2008). The difficulties were further amplified in nonlinear models, where sums and products are no longer applicable. As demands grew to tackle problems involving categorical variables and nonlinear interactions, researchers could no longer define direct and indirect effects in terms of structural or regressional coefficients, and all attempts to extend the linear paradigms of effect decomposition to nonlinear systems produced distorted results (MacKinnon et al., 2007). The counterfactual reading of structural equations (6) enables us to redefine and analyze direct and indirect effects from first principles, uncommitted to distributional assumptions or a particular parametric form of the equations.

5.4.2 Direct Effects

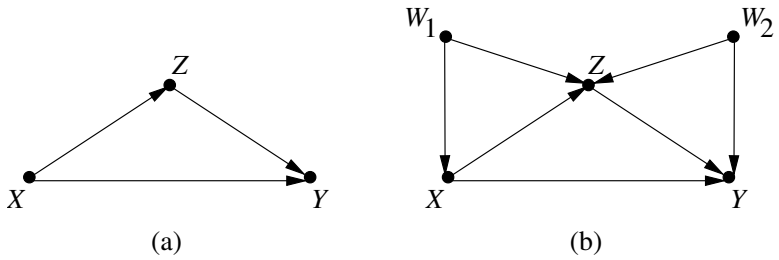

Figure 8: A generic model depicting mediation through Z (a) with no confounders and (b) with two confounders, W_1 and W_2.

Conceptually, we can define the direct effect $DE_{x,x'}(Y)$[11] as the expected change in Y induced by changing X from x to x' while keeping all mediating factors constant at whatever value they *would have obtained* under $do(x)$ (Robins and Greenland, 1992; Pearl, 2001). Accordingly, Pearl (2001) defined direct effect using counterfactual notation:

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \tag{9}$$

Here, Y_{x',Z_x} represents the value that Y would attain under the operation of setting X to x' and, simultaneously, setting Z to whatever value it would have obtained under the setting $X = x$. Given certain assumptions of "no confounding," it is possible to show Pearl (2001) that the direct effect can be reduced to a *do*-expression:

$$DE_{x,x'}(Y) = \sum_{zw} [E(Y|(do(x', z), w) - E(Y|do(x, z), w)]P(z|do(x), w)P(w) \tag{10}$$

where W satisfies the back-door criterion relative to both $X \to Z$ and $(X, Z) \to Y$.

[11]Robins and Greenland (1992) called this notion of direct effect "Pure" while Pearl called it "Natural," denoted NDE, to be distinguished from the "controlled direct effect" which is specific to one level of the mediator Z. We will delete the letter "N" from the acronyms of both the direct and indirect effect, and use DE and IE, respectively.

In particular, expression (10) is both valid and identifiable in Markovian models (i.e., no unobserved confounders) where each term on the right can be reduced to a "*do*-free" expression using equation (8) and then estimated by regression.

For example, for the model in Figure 8(b), equation (10) reads

$$DE_{x,x'}(Y) = \sum_z \sum_{w_2} P(w_2)[E(Y|x', z, w_2)) - E(Y|x, z, w_2))] \sum_{w_1} P(z|x, w_1, w_2)P(w_1). \quad (11)$$

while for the confounding-free model of Figure 8(a) we have

$$DE_{x,x'}(Y) = \sum_z [E(Y|x', z) - E(Y|x, z)]P(z|x). \quad (12)$$

Both (11) and (12) can be estimated by a two-step regression.

5.5 Indirect Effects

Remarkably, the definition of the direct effect (9) can be turned around and provide an operational definition for the *indirect effect*—a concept shrouded in mystery and controversy, because it is impossible, by controlling any of the variables in the model, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

The *indirect effect*, IE, of the transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z to whatever value it would have attained had X been set to $X = x'$. Formally, this reads

$$IE_{x,x'}(Y) \stackrel{\Delta}{=} E[(Y_{x,Z_{x'}}) - E(Y_x)], \quad (13)$$

which is almost identical to the direct effect (equation 9) save for exchanging x and x' in the first term (Pearl, 2001).

Indeed, it can be shown that, in general, the total effect TE of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the reverse transition. Formally,

$$TE_{x,x'}(Y) \stackrel{\Delta}{=} E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \quad (14)$$

In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \quad (15)$$

Since each term above is based on an independent operational definition, this equality constitutes a formal justification for the additive formula used routinely in linear systems.

5.6 The Mediation Formula: A Simple Solution to a Thorny Problem

This subsection demonstrates how the solution provided in equations (12) and (15) can be applied in assessing mediation effects in nonlinear models. We will use the simple mediation

model of Figure 8(a), where all error terms (not shown explicitly) are assumed to be mutually independent, with the understanding that adjustment for appropriate sets of covariates W may be necessary to achieve this independence (as in equation 11) and that integrals should replace summations when dealing with continuous variables (Imai et al., 2010).

Combining (12) and (14), the expression for the indirect effect, IE, becomes

$$IE_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)] \tag{16}$$

which provides a general formula for mediation effects, applicable to any nonlinear system, any distribution (of U), and any type of variables. Moreover, the formula is readily estimable by regression. Owing to its generality and ubiquity, I have referred to this expression as the "Mediation Formula" (Pearl, 2009, 2010b).

The Mediation Formula represents the average increase in the outcome Y that the transition from $X = x$ to $X = x'$ is expected to produce absent any direct effect of X on Y. Though based on solid causal principles, it embodies no causal assumption other than the generic mediation structure of Figure 8(a). When the outcome Y is binary (e.g., recovery, or hiring) the ratio $(1 - IE/TE)$ represents the fraction of responding individuals who owe their response to direct paths, while $(1 - DE/TE)$ represents the fraction who owe their response to Z-mediated paths.

The Mediation Formula tells us that IE depends only on the expectation of the counterfactual Y_{xz}, not on its functional form $f_Y(x, z, u_Y)$ or its distribution $P(Y_{xz} = y)$. It calls therefore for a two-step regression which, in principle, can be performed nonparametrically. In the first step we regress Y on X and Z, and obtain the estimate

$$g(x,z) = E(Y|x,z) \tag{17}$$

for every (x, z) cell. In the second step we fix x and regard $g(x, z)$ as a function $g_x(z)$ of Z. We now estimate the conditional expectation of $g_x(z)$, conditional on $X = x'$ and $X = x$, respectively, and take the difference

$$IE_{x,x'}(Y) = E_{Z|X}[g_x(z)|x'] - E_{Z|X}[g_x(z)|x']. \tag{18}$$

Nonparametric estimation is not always practical. When Z consists of a vector of several mediators, the dimensionality of the problem might prohibit the estimation of $E(Y|x, z)$ for every (x, z) cell, and the need arises to use parametric approximation. We can then choose any convenient parametric form for $E(Y|x, z)$ (e.g., linear, logit, probit), estimate the parameters separately (e.g., by regression or maximum likelihood methods), insert the parametric approximation into (16) and estimate its two conditional expectations (over z) to get the mediated effect (VanderWeele, 2009).

Let us examine what the Mediation Formula yields when applied to the linear version of Figure 8(a), which reads

$$\begin{aligned} x &= u_X \\ z &= b_0 + \beta x + u_Z \\ y &= c_0 + \alpha x + \gamma z + u_Y \end{aligned} \tag{19}$$

with u_X, u_Y, and u_Z uncorrelated, zero-mean error terms. Computing the conditional expectation in (16) gives

$$E(Y|x,z) = E(c_0 + \alpha x + \gamma z + u_Y) = c_0 + \alpha x + \gamma z$$

and yields

$$\begin{aligned} IE_{x,x'}(Y) &= \sum_z (\alpha x + \gamma z)[P(z|x') - P(z|x)]. \\ &= \gamma[E(Z|x') - E(Z|x)] \qquad (20) \\ &= (x' - x)(\beta\gamma) \qquad (21) \\ &= (x' - x)(\tau - \alpha) \qquad (22) \end{aligned}$$

where τ is the slope of the total effect;

$$\tau = (E(Y|x') - E(Y|x))/(x' - x) = \alpha + \beta\gamma.$$

We thus obtained the standard expressions for indirect effects in linear systems, which can be estimated either as a difference $\tau - \alpha$ of two regression coefficients (equation 22) or as a product $\beta\gamma$ of two regression coefficients (equation 21) (see MacKinnon et al., 2007). These two strategies do not generalize to nonlinear systems; direct application of (16) is necessary Pearl (2010a).

To understand the difficulty, assume that the correct model behind the data contains a product term δxz added to equation (19), giving:

$$y = c_0 + \alpha x + \gamma z + \delta xz + u_Y$$

Further assume that we correctly account for this added term and, through sophisticated regression analysis, we obtain accurate estimates of all parameters in the model. It is still not clear what combinations of parameters measure the direct and indirect effects of X on Y, or, more specifically, how to assess the fraction of the total effect that is *explained* by mediation and the fraction that is *owed* to mediation. In linear analysis, the former fraction is captured by the product $\beta\gamma/\tau$ (equation 21), the latter by the difference $(\tau - \alpha)/\tau$ (equation 22) and the two quantities coincide. In the presence of interaction, however, each fraction demands a separate analysis, as dictated by the Mediation Formula.

To witness, substituting the nonlinear equation in (12), (15) and (16) and assuming $x = 0$ and $x' = 1$, yields the following effect decomposition:

$$\begin{aligned} DE &= \alpha + b_0\delta \\ IE &= \beta\gamma \\ TE &= \alpha + b_0\delta + \beta(\gamma + \delta) \\ &= DE + IE + \beta\gamma \end{aligned}$$

We therefore conclude that the portion of output change for which mediation would be *sufficient* is

$$IE = \beta\gamma$$

while the portion for which mediation would be *necessary* is

$$TE - DE = \beta\gamma + \beta\delta$$

We note that, due to interaction, a direct effect can be sustained even when the parameter α vanishes and, moreover, a total effect can be sustained even when both the direct and indirect effects vanish. This illustrates that estimating parameters in isolation tells us little about the effect of mediation and, more generally, mediation and moderation are intertwined and cannot be assessed separately.

If the policy evaluated aims to prevent the outcome Y by ways of weakening the mediating pathways, the target of analysis should be the difference $TE - DE$, which measures the highest prevention potential of any such policy. If, on the other hand, the policy aims to prevent the outcome by weakening the direct pathway, the target of analysis should shift to IE, for $TE - IE$ measures the highest preventive potential of this type of policies.

The main power of the Mediation Formula shines in studies involving categorical variables, especially when we have no parametric model of the data generating process. To illustrate, consider the case where all variables are binary, still allowing for arbitrary interactions and arbitrary distributions of all processes. The low dimensionality of the binary case permits both a nonparametric solution and an explicit demonstration of how mediation can be estimated directly from the data. Generalizations to multi-valued outcomes are straightforward.

Assume that the model of Figure 8(a) is valid and that the observed data is given by Table 3. The factors $E(Y|x,z)$ and $P(Z|x)$ can be readily estimated as shown in the two

Number of Samples	X	Z	Y	$E(Y\|x,z) = \boldsymbol{g_{xz}}$	$E(Z\|x) = \boldsymbol{h_x}$
n_1	0	0	0	$\frac{n_2}{n_1+n_2} = g_{00}$	$\frac{n_3+n_4}{n_1+n_2+n_3+n_4} = h_0$
n_2	0	0	1		
n_3	0	1	0	$\frac{n_4}{n_3+n_4} = g_{01}$	
n_4	0	1	1		
n_5	1	0	0	$\frac{n_6}{n_5+n_6} = g_{10}$	$\frac{n_7+n_8}{n_5+n_6+n_7+n_8} = h_1$
n_6	1	0	1		
n_7	1	1	0	$\frac{n_8}{n_7+n_8} = g_{11}$	
n_8	1	1	1		

Table 3: Computing the Mediation Formula for the model in Figure 8(a), with X, Y, Z binary.

right-most columns of Table 3 and, when substituted in (12), (15), (16), yield

$$DE = (g_{10} - g_{00})(1 - h_0) + (g_{11} - g_{01})h_0 \tag{23}$$

$$IE = (h_1 - h_0)(g_{01} - g_{00}) \tag{24}$$

$$TE = g_{11}h_1 + g_{10}(1 - h_1) - [g_{01}h_0 + g_{00}(1 - h_0)] \tag{25}$$

We see that logistic or probit regression is not necessary; simple arithmetic operations suffice to provide a general solution for any conceivable data set, regardless of the data-generating process.

Numerical example

To anchor these formulas in a concrete example, let us assume that $X = 1$ stands for a drug treatment, $Y = 1$ for recovery, and $Z = 1$ for the presence of a certain enzyme in a patient's blood which appears to be stimulated by the treatment. Assume further that the data described in Tables 4 and 5 was obtained in a randomized clinical trial and that our research question is whether Z mediates the action of X on Y, or is merely a catalyst that accelerates the action of X on Y.

Treatment X	Enzyme present Z	Percentage cured $g_{xz} = E(Y\|x, z)$
YES	YES	$g_{11} = 80\%$
YES	NO	$g_{10} = 40\%$
NO	YES	$g_{01} = 30\%$
NO	NO	$g_{00} = 20\%$

Table 4:

Treatment X	Percentage with Z present
NO	$h_0 = 40\%$
YES	$h_1 = 75\%$

Table 5:

Substituting this data into Eqs. (23)–(25) yields:

$$DE = (0.40 - 0.20)(1 - 0.40) + (0.80 - 0.30)0.40 = 0.32$$
$$IE = (0.75 - 0.40)(0.30 - 0.20) = 0.035$$
$$TE = 0.80 \times 0.75 + 0.40 \times 0.25 - (0.30 \times 0.40 + 0.20 \times 0.10) = 0.46$$
$$IE/TE = 0.07 \qquad DE/TE = 0.696 \qquad 1 - DE/TE = 0.304$$

We conclude that 30.4% of those recovered owe their recovery to the capacity of the treatment to stimulate the secretion of the enzyme, while only 7% of recoveries would be sustained by enzyme stimulation alone. The enzyme seems to act more as a catalyst for the healing process of X than having a healing action of its own. The policy implication of such a study would be that efforts to substitute the drug with an alternative stimulant of the enzyme are not likely to be effective, the drug evidently has a direct effect on the disease, independent of, albeit enhanced by enzyme stimulation.

In comparing these results to those produced by conventional mediation analyses we should note that conventional methods do not define direct and indirect effects in a setting where the underlying process is unknown. MacKinnon (2008, Ch. 11), for example, analyzes categorical data using logistic and probit regressions and constructs effect measures using products and differences of the parameters in those regressional forms. This strategy is not

compatible with the causal interpretation of effect measures, even when the parameters are precisely known; IE and DE may be extremely complicated functions of those regression coefficients (Pearl, 2010b). Fortunately, those coefficients need not be estimated at all; effect measures can be estimated directly from the data, circumventing the parametric analysis altogether, as shown in equations (23)–(25).

Attempts to extend the difference and product heuristics to nonparametric analysis have encountered ambiguities that conventional analysis fails to resolve.

The product-of-coefficients heuristic advises us to multiply the unit effect of X on Z

$$C_\beta = E(Z|X = 1) - E(Z|X = 0) = h_1 - h_0$$

by the unit effect of Z on Y given X,

$$C_\gamma = E(Y|X = x, Z = 1) - E(Y|X = x, Z = 0) = g_{x1} - g_{x0}$$

but does not specify on what value we should condition X. Equation (24) resolves this ambiguity by determining that C_γ should be conditioned on $X = 0$; only then would the product $C_\beta C_\gamma$ yield the correct mediation measure, IE.

The difference-in-coefficients heuristics instructs us to estimate the direct effect coefficient

$$C_\alpha = E(Y|X = 1, Z = z) - E(Y|X = 0, Z = z) = g_{1z} - g_{0z}$$

and subtract it from the total effect, but does not specify on what value we should condition Z. Equation (23) determines that the correct way of estimating C_α would be to condition on both $Z = 0$ and $Z = 1$ and take their weighted average, with $h_0 = P(Z = 1|X = 0)$ serving as the weighting function.

To summarize, the Mediation Formula dictates that, in calculating IE, we should condition on both $Z = 1$ and $Z = 0$ and average while, in calculating DE, we should condition on only one value, $X = 0$, and no average need be taken.

The difference and product heuristics are both legitimate, with each seeking a different effect measure. The difference heuristics, leading to $TE - DE$, seeks to measure the percentage of units for which mediation was *necessary*. The product heuristics on the other hand, leading to IE, seeks to estimate the percentage of units for which mediation was *sufficient*. The former informs policies aiming to modify the direct pathway while the latter informs those aiming to modify mediating pathways.

In addition to providing causally sound estimates for mediation effects, the Mediation Formula also enables researchers to evaluate analytically the effectiveness of various parametric specifications relative to any assumed model. This type of analytical "sensitivity analysis" has been used extensively in statistics for parameter estimation but could not be applied to mediation analysis, owing to the absence of an objective target quantity that captures the notion of indirect effect in both linear and nonlinear systems, free of parametric assumptions. The Mediation Formula of equation (16) explicates this target quantity formally, and casts it in terms of estimable quantities. It has also been used by Imai et al. (2010) to examine the robustness of empirical findings to the possible existence of unmeasured confounders.

The derivation of the Mediation Formula was facilitated by taking seriously the graphical-counterfactual-structural symbiosis spawned by the surgical interpretation of counterfactuals

(equation 6). In contrast, when the mediation problem is approached from an exclusivist potential-outcome viewpoint, void of the structural guidance of equation (6), counterintuitive definitions ensue, carrying the label "principal stratification" (Rubin, 2004, 2005), which are at variance with common understanding of direct and indirect effects. For example, the direct effect is definable only in units absent of indirect effects. This means that a grandfather would be deemed to have no direct effect on his grandson's behavior in families where he has had some effect on the father. This precludes from the analysis all typical families, in which a father and a grandfather have simultaneous, complementary influences on children's upbringing. In linear systems, to take an even sharper example, the direct effect would be undefined whenever indirect paths exist from the cause to its effect. The emergence of such paradoxical conclusions underscores the wisdom, if not necessity of a symbiotic analysis, in which the counterfactual notation $Y_x(u)$ is governed by its structural definition, equation (6).[12]

6 Conclusions

This chapter casts the methodology of structural equation modeling as a causal-inference engine that takes qualitative causal assumptions, data and queries as inputs and produces quantitative causal claims, conditional on the input assumptions, together with data-fitness ratings to well-defined statistical tests.

We show that graphical encodings of the input assumption can also be used as efficient mathematical tools for identifying testable implications, deciding query identification and generating estimable expressions for causal and counterfactual expressions. We discussed the logical equivalence of the structural and potential-outcome frameworks and demonstrated the advantages of a symbiotic approach by offering a simple solution to the mediation problem for models with categorical data.

An issue that was not discussed in this chapter is the perennial problem of external validity (Shadish et al., 2002), namely, the conditions under which causal conclusions from a study on one population can safely be modified and transported to another. This problem has recently received a formal treatment using nonparametric SEM, and has led to algorithmic criteria for deciding the legitimacy of the transfer as well as the way it ought to be executed (Pearl and Bareinboim, 2010).

Some researchers would naturally prefer a methodology in which claims are less sensitive to judgmental assumptions; unfortunately, no such methodology exists. The relationship between assumptions and claims is a universal one—namely, for every set A of assumptions (knowledge) there is a unique set of conclusions C that one can deduce from A, given the data, regardless of the method used. The completeness results of Shpitser and Pearl (2006a) imply that SEM operates at the boundary of this universal relationship; no method can do better without strengthening the assumptions.

[12]Such symbiosis is now standard in epidemiology research (Robins, 2001; Petersen et al., 2006; VanderWeele and Robins, 2007; Hafeman and Schwartz, 2009; Joffe and Green, 2009; VanderWeele, 2009; Kaufman, 2010) and is making its way slowly toward the social and behavioral sciences (Morgan and Winship, 2007; Imai et al., 2010).

Acknowledgments

This paper has benefited from discussions with Elias Barenboim, Peter Bentler, Ken Bollen, James Heckman, Jeffrey Hoyle, Marshal Joffe, David Kaplan, David Kenny, David MacKinnon, Rod McDonald, William Shadish, and Leland Wilkinson, and was supported in parts by grants from NIH #1R01 LM009961-01, NSF #IIS-0914211 and #IIS-1018922, and ONR #N000-14-09-1-0665.

References

Ali, R., Richardson, T. and Spirtes, P. (2009). Markov equivalence for ancestral graphs. *The Annals of Statistics* **37** 2808–2837.

Balke, A. and Pearl, J. (1995). Counterfactuals and policy analysis in structural models. In *Uncertainty in Artificial Intelligence, Proceedings of the Eleventh Conference* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 11–18.

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

Baumrind, D. (1993). Specious causal attributions in social sciences: The reformulated stepping-stone theory of hero in use as exemplar. *Journal of Personality and Social Psychology* **45** 1289–1298.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* **2** 47–53.

Blalock, H. (1964). *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill.

Bollen, K. (1989). *Structural Equations with Latent Variables*. John Wiley, New York.

Brito, C. and Pearl, J. (2002). Generalized instrumental variables. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference* (A. Darwiche and N. Friedman, eds.). Morgan Kaufmann, San Francisco, 85–93.

Byrne, B. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming*. 2nd ed. Routledge, New York.

Cai, Z. and Kuroki, M. (2008). On identifying total effects in the presence of latent variables and selection bias. In *Uncertainty in Artificial Intelligence, Proceedings of the Twenty-Fourth Conference* (D. McAllester and P. Myllymäki, eds.). AUAI, Arlington, VA, 62–69.

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, New York, NY.

Chin, W. (1998). Commentary: Issues and opinion on structural equation modeling. *Management Information Systems Quarterly* **22** 7–16.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research* **18** 115–126.

Duncan, O. (1975). *Introduction to Structural Equation Models*. Academic Press, New York.

Freedman, D. (1987). As others see us: A case study in path analysis (with discussion). *Journal of Educational Statistics* **12** 101–223.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science* **3** 151–182.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

Hafeman, D. and Schwartz, S. (2009). Opening the black box: A motivation for the assessment of mediation. *International Journal of Epidemiology* **3** 838–845.

Halpern, J. (1998). Axiomatizing causal reasoning. In *Uncertainty in Artificial Intelligence* (G. Cooper and S. Moral, eds.). Morgan Kaufmann, San Francisco, CA, 202–210. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

Heckman, J. (2005). The scientific model of causality. *Sociological Methodology* **35** 1–97.

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

Holland, P. (1988). Causal inference, path analysis, and recursive structural equations models. In *Sociological Methodology* (C. Clogg, ed.). American Sociological Association, Washington, D.C., 449–484.

Holland, P. (1995). Some reflections on Freedman's critiques. *Foundations of Science* **1** 50–57.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* **25** 51–71.

Imbens, G. (2010). An economists perspective on Shadish (2010) and West and Thoemmes (2010). *Psychological Methods* **15** 47–55.

Joffe, M. and Green, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* 530–538.

Kaufman, J. (2010). Invited commentary: Decomposing with a lot of supposing. *American Journal of Epidemiology* **172** 1349–1351.

Kelloway, E. (1998). *Using LISREL for structural Equation Modeling*. Sage, Thousand Oaks, CA.

Kline, R. (2005). *Principles and Practice of Structural Equation Modeling*. 2nd ed. The Guilford Press, New York.

Koopmans, T. (1953). Identification problems in econometric model construction. In *Studies in Econometric Method* (W. Hood and T. Koopmans, eds.). Wiley, New York, 27–48.

Kyono, T. (2010). Commentator: A front-end user-interface module for graphical and structural equation modeling. Tech. Rep. R-364, <http://ftp.cs.ucla.edu/pub/stat_ser/r364.pdf>, Master Thesis, Department of Computer Science, University of California, Los Angeles, CA.

Lee, S. and Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research* **25** 313–334.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates, New York.

MacKinnon, D., Lockwood, C., Brown, C., Wang, W. and Hoffman, J. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials* **4** 499–513.

Marschak, J. (1950). Statistical inference in economics. In *Statistical Inference in Dynamic Economic Models* (T. Koopmans, ed.). Wiley, New York, 1–50. Cowles Commission for Research in Economics, Monograph 10.

McDonald, R. (2002). What can we learn from the path equations?: Identifiability constraints, equivalence. *Psychometrika* **67** 225–249.

McDonald, R. (2004). The specific analysis of structural equation models. *Multivariate Behavioral Research* **39** 687–713.

Morgan, S. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. Cambridge University Press, New York, NY.

Muthén, B. (1987). Response to Freedman's critique of path analysis: Improve credibility by better methodological training. *Journal of Educational Statistics* **12** 178–184.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science* **8** 266–269.

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

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. (2001). Direct and indirect effects. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference* (J. Breese and D. Koller, eds.). Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, J. (2004). Robustness of causal claims. In *Proceedings of the Twentieth Conference Uncertainty in Artificial Intelligence* (M. Chickering and J. Halpern, eds.). AUAI Press, Arlington, VA, 446–453.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

Pearl, J. (2010a). An introduction to causal inference. *The International Journal of Biostatistics* **6** DOI: 10.2202/1557–4679.1203, <http://www.bepress.com/ijb/vol6/iss2/7/>.

Pearl, J. (2010b). The mediation formula: A guide to the assessment of causal pathways in non-linear models. Tech. Rep. R-363, <http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf>, Department of Computer Science, University of California, Los Angeles, CA. To appear in C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.), *Statistical Causality*. Forthcoming, 2011.

Pearl, J. (2010c). On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (P. Grünwald and P. Spirtes, eds.). AUAI, Corvallis, OR, 425–432. <http://ftp.cs.ucla.edu/pub/stat_ser/r357.pdf>.

Pearl, J. and Bareinboim, E. (2010). Transportability across studies: A formal approach. Tech. Rep. R-372, <http://ftp.cs.ucla.edu/pub/stat_ser/r372.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Submitted to JSM.

Petersen, M., Sinisi, S. and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology* **17** 276–284.

Robins, J. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12** 313–320.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

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

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

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

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

Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 437–444.

Shpitser, I. and Pearl, J. (2006b). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 1219–1226.

Shpitser, I. and Pearl, J. (2008). Dormant independence. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 1081–1087.

Simon, H. (1953). Causal ordering and identifiability. In *Studies in Econometric Method* (W. C. Hood and T. Koopmans, eds.). Wiley and Sons, Inc., New York, NY, 49–74.

Simon, H. and Rescher, N. (1966). Cause and counterfactual. *Philosophy and Science* **33** 323–340.

Sobel, M. (1996). An introduction to causal inference. *Sociological Methods & Research* **24** 353–379.

Sobel, M. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* **33** 230–231.

Sørensen, A. (1998). Theoretical methanisms and the empirical study of social processes. In *Social Mechanisms: An Analytical Approach to Social Theory, Studies in Rationality and Social Change* (P. Hedström and R. Swedberg, eds.). Cambridge University Press, Cambridge, 238–266.

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research* **21** 309–331.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence* (R. Dechter, M. Kearns and R. Sutton, eds.). AAAI Press/The MIT Press, Menlo Park, CA, 567–573.

VanderWeele, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20** 18–26.

VanderWeele, T. and Robins, J. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology* **18** 561–568.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence, Proceedings of the Sixth Conference*. Cambridge, MA. Also in P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 6*, Elsevier Science Publishers, B.V., 255–268, 1991.

West, S. and Thoemmes, F. (2010). Campbells and Rubins perspectives on causal inference. *Psychological Methods* **15** 18–37.

Wilkinson, L., the Task Force on Statistical Inference and *APA Board of Scientific Affairs* (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* **54** 594–604.

Wright, P. (1928). *The Tariff on Animal and Vegetable Oils*. The MacMillan Company, New York, NY.

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

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