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

July 29, 2010

*Portions of this paper are adapted from Pearl (2000a, 2009b,a, 2010a); I am indebted to Peter Bentler, Felix Elwert, David MacKinnon, Stephen Morgan, Patrick Shrout, Christopher Winship, and many readers of the UCLA Causality Blog (http://www.mii.ucla.edu/causality/) for reading and commenting on various segments of this manuscript, and to two anonymous referees for their thorough editorial input. This research was supported in parts by NIH grant #1R01 LM009961-01, NSF grant #IIS-0914211, and ONR grant #N000-14-09-1-0665.

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 an astute warning: "It is sometimes thought that correlation does not prove causation but 'causal modeling' does... [McDonald] showed the dangers of this practice." But ends with a surprising 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, and all modern advances in graphical and non-parametric structural models have all been misguided, for they have been chasing parameters that have no causal interpretation.

The absurdity of such conclusions 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 manipulative experiments, and "interpreting parameters as causal effects," which

may be based on scientific knowledge. One can legitimately be in a possession of a parameter that stands for a causal effect and still be unable, using statistical means, 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 a parameter with such interpretation. Science is not prepared to accept this defeatist verdict.

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, and describe the tools needed for solving most (if not all) problems involving causal relationships. The tools are based on non-parametric 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 and illucidate it causal content. SEM from its parametric needs edit blinders.

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 to 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 non-experimental 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

8. A simple, causally-based solution to the so called "Mediation Problem," (Baron and Kenny, 1986; MacKinnon, 2008). It takes the form of formulas for direct and indirect effects that are applicable to both continuous and categorical variables, linear and nonlinear interactions, and are readily estimable by regression.

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 proper 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 sociological 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 (2009b, p. 138) attributes to notational deficiency; 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 identify his error and to articulate the correct, noncontradictory interpretation of the example presented by Freedman. Instead, SEM researchers appeared willing to accept the contradiction as a fundamental flaw in causal thinking, which must always give way to statistical superiority. In his highly cited commentary on SEM, Chin (1998) surrenders to the critic: "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 approach to causal inference (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, careful logical 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 3).

Paul Holland, another advocate of the potential-outcome framework, unveiled the source 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 $\{y\}$ (see footnote **??**).

We will further see (Section **??**) 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 non-parametric perspective unveils their potentials and avails them for both linear and nonlinear analyses.

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

3 The logic of SEM (via email)

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

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

There is no data that could possibly refute this statement so, goodness-of-fit with any data would not uncover inconsistency in the hypothesized model and, still, 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

$M =$ "The rooster causal the sun to rise and the average age in Los Angeles is higher than 3"

will encounter a similar objection; although M is now falsifiable, its success in fitting the data tells us nothing about Rooster crows.

This impediment is not a contrived caricature of SEM, or of the confirmatory approach;

it represents however a profound logical flaws of any approach that does not spell out the empirical content of the assumptions behind the hypothesized model and the claims it makes upon clearing the scrutiny of the data.

The interpretation of SEM methodology that emerges from the non-parametric perspective (Pearl, 2009b, 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 two inputs and produces two outputs. The inputs are:

I-1. A model M that encodes a set A of qualitative causal assumptions which the investigator is prepared to defend on scientific grounds. (A is typically encoded in 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 D of experimental or non-experimental data, presumably generated by a process consistent with A.

The outputs are

O-1. A set C of quantitative claims about the magnitudes of causal and counterfactual relationships among variables of interest, conditional on A. A typical claim would be that the causal influence of one variable on another (represented by path coeffcients in linear exceeds a certain amount, or is mediated substantially by a third variable. Whether causal relationships of this sort can be ascertained consistently from D depends on their *identifiability* – a property that can be decided from M prior to taking

my data.

O-2. A list T of testable statistical implications of A, and the degree to which the data refutes each of those implications. A typical implication would be the vanishing of a regression coefficient, or other constraints on the covariance matrix, which, again, can be determined from M prior to taking any data.

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

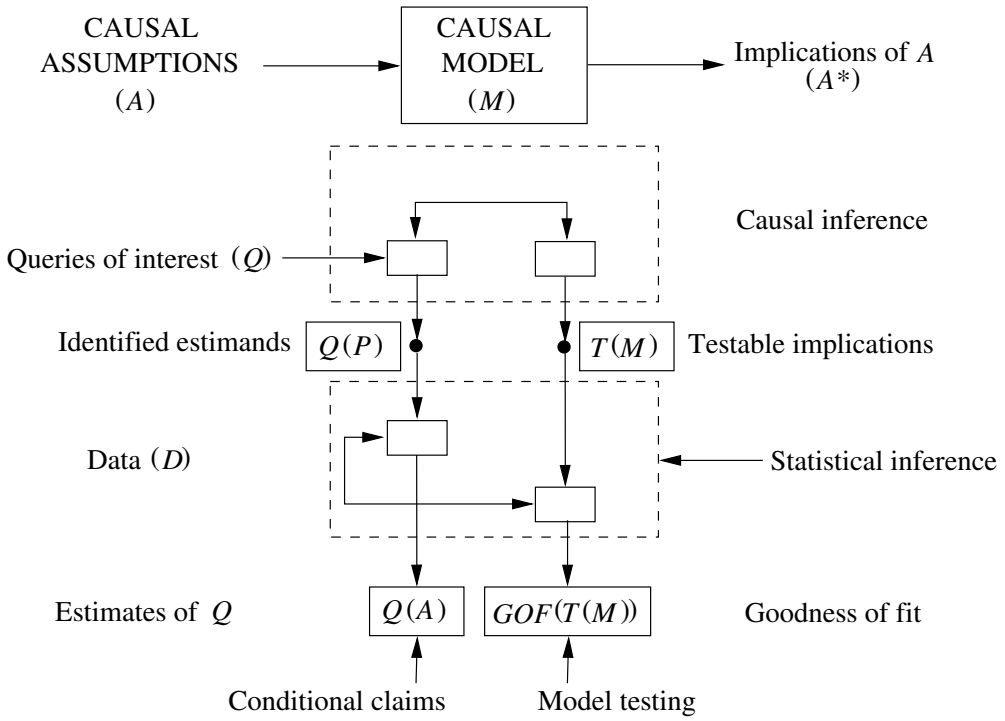

Figure 1: atlanta slide1

Several observations are worth noting before illustrating these inferences by examples. First, SEM is not a statistical methodology; its aims cannot be described in terms of statistical tasks, be it hypothesis testing or estimation, since causal and counterfactual claims cannot be defined in terms of density functions of realizable variables.

Second, all claims produced by an SEM study are conditional on the validity of A, and should be reported in implicative format: "If A then C_i" for any claim $C_i \in C$ that is of interest. Yet, despite their conditional part, such claims 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 (GOF) test is not a necessary prerequisite for the logical validity of the conditional claim "If A then C_i," nor is it necessary for the empirical validity of C_i. While it is important to know if any assumptions in A are inconsistent with the data, M may not have any testable implications whatsoever and, still, the assertion "If A then C_i" may be extremely informative in a decision making context, for C_i commits to quantitative relationships rather then the 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 models 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 goodness-of-fit measures hide this information.

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,

[2]It is important to emphasize this point 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 require the same knowledge, though some tend to hide the assumptions under technical verbiage such is "ignorability" or "nonconfoundedness."

namely, models that satisfy all the testable implications of M and still advertise claims that contradict C. Modern developments in graphical modeling have devised visual and algorithmic tools for detecting, displaying, and enumerating of such models. Researches researchers? should keep in mind though that only a tiny portion of the assumptions behind each SEM study lends itself to statistical tests; the bulk of it must remain untestable, at the mercy of scientific judgment.

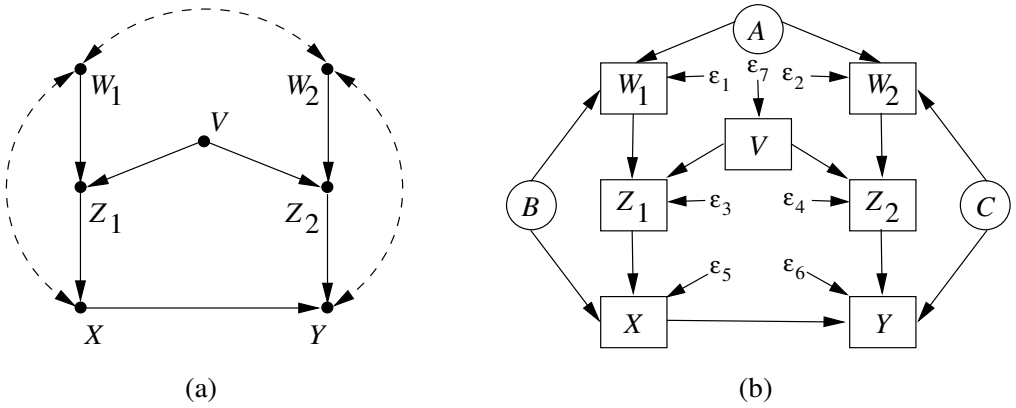

Figure 2: (a) A Sample Path Diagram Pearl (2009b). (Note: latent variables are not shown explicitly, but are presumed to emit the double-head arrows.) (b) Conventional Path Diagram (Note: identical to (a) with latent variables and errors shown explicitly).

4 Section 4 - r360

1. 360: section 3, section 4
2. R 364: Figure 1 a and b, section 2
3. R 355: Section 3.1, section 4.3.1, section 6

4.1 An Outline of the Structural Theory (sec3-r360)

The analysis illustrated in the preceding section is part of a general theory of counterfactuals that I named "structural" (Pearl, 2000a, Chapter 7) in honor of its origin in the structural equation models developed by econometricians in the 1940-50's (Haavelmo, 1943; Simon, 1953; Hurwicz, 1950; Marschak, 1953).

At the center of the theory lies a "structural model," M, consisting of two sets of variables, U and V, and a set F of functions that determine how values are assigned to each variable $V_i \in V$. Thus for example, the equation

$$v_i = f_i(v, u)$$

describes a physical process by which Nature *examines* the current values, v and u, of all variables in V and U and, accordingly, *assigns* variable V_i the value $v_i = f_i(v, u)$. The variables in U are considered "exogenous," namely, background conditions for which no explanatory mechanism is encoded in model M. Every instantiation $U = u$ of the exogenous variables uniquely determines the values of all variables in V and, hence, if we assign a probability $P(u)$ to U, it defines a probability function $P(v)$ on V.

The basic counterfactual entity in structural models is the sentence: "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$. The key to interpreting counterfactuals is to treat the subjunctive phrase "had X been x" as an instruction to make a minimal modification in the current model, so as to ensure the antecedent condition $X = x$. Such a minimal modification amounts to replacing the equation for X by a constant x, as we have done in Fig. **??**(c). This replacement permits the constant x to differ from the actual value of X (namely $f_x(v, u)$) without rendering the system of equations inconsistent, thus allowing

all variables, exogenous as well as endogenous, to serve as antecedents.

Letting M_x stand for a modified version of M, with the equation(s) of X replaced by $X = x$, the formal definition of the counterfactual $Y_x(u)$ reads:

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{1}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x. Galles and Pearl (1998) and Halpern (1998) have given a complete axiomatization of structural counterfactuals, embracing both recursive and non-recursive models. (see also Pearl, 2009b, Chapter 7).

Since the distribution $P(u)$ induces a well defined probability on the counterfactual event $Y_x = y$, it also defines a joint distribution on all Boolean combinations of such events, for instance '$Y_x = y$ AND $Z_{x'} = z$,' which may appear contradictory, if $x \neq x'$. For example, to answer retrospective questions, such as whether Y would be y_1 if X were x_1, given that in fact Y is y_0 and X is x_0, we need to compute the conditional probability $P(Y_{x_1} = y_1|Y = y_0, X = x_0)$ which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model.

In general, the probability of the counterfactual sentence $P(Y_x = y|e)$, where e is any propositioned evidence, can be computed by the 3-step process (illustrated in Section 2);

Step 1 (***abduction***): Update the probability $P(u)$ to obtain $P(u|e)$.

Step 2 (***action***): Replace the equations corresponding to variables in set X by the equations $X = x$.

Step 3 (***prediction***): Use the modified model to compute the probability of $Y = y$.

In temporal metaphors, Step 1 explains the past (U) in light of the current evidence e; Step 2 bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$. It can be shown (Pearl, 2000a, p. 76) that this procedure can be given an interpretation in terms of "imaging" (Lewis, 1973) – a process of "mass-shifting" among possible worlds – provided that (a) worlds with equal histories should be considered equally similar and (b) equally-similar worlds should receive mass in proportion to their prior probabilities (Pearl, 2000a, pp. 76).

4.2 Summary of Applications (sec4-r360)

Since its inception (Balke and Pearl, 1995) this counterfactual model has provided mathematical solutions to a number of problems in policy analysis and retrospective reasoning. In the context of decision making, for example, a rational agent is instructed to maximize the expected utility

$$EU(x) = \sum P(Y_x = y)U(y) \tag{2}$$

over all options x. Here, $U(y)$ stands for the utility of outcome $Y = y$ and $P(Y_x = y)$ stands for the probability that outcome $Y = y$ would prevail, had action $do(X = x)$ been performed and condition $X = x$ firmly established.[3]

The central question in many of the empirical sciences is that of *identification*: Can we predict the effect of a contemplated action $do(X = x)$ or, in other words, can the

[3]Equation (2) represents the dictates of Causal Decision Theory (CDT) Stalnaker (1972); Lewis (1973); Gardenfors (1988) and Joyce (1999) – the pitfalls of Evidential Decision Theory are well documented (see (Skyrms, 1980; Pearl, 2000a, pp. 108–9)), and need not be considered.

post-intervention distribution, $P(Y_x = y)$, be estimated from data generated by the pre-intervention distribution, $P(z, x, y)$? Clearly, since the prospective counterfactual Y_x is generally not observed, the answer must depend on the agent's model M and then the question reduces to: Can $P(Y_x = y)$ be estimated from a combination of $P(z, x, y)$ and a graph G that encodes the structure of M.

This problem has been solved by deriving a precise characterization of what Skyrms (1980) called "K-partition," namely, a set S of observed variables that permits $P(Y_x = y)$ to be written in terms of Bayes conditioning on, or, "adjusting for" S:

$$P(Y_x = y) = \sum_s P(y|x, s)P(s)$$

Tian and Pearl (2002) and Shpitser and Pearl (2007) further expanded this result and established a criterion that permits (or forbids) the assessment of $P(Y_x = y)$ by any method whatsoever.

Prospective counterfactual expressions of the type $P(Y_x = y)$ are concerned with predicting the average effect of hypothetical actions and policies and can, in principle, be assessed from experimental studies in which X is randomized. Retrospective counterfactuals, on the other hand, like S_2 in the Oswald scenario, consist of variables at different hypothetical worlds (different subscripts) and these may or may not be testable experimentally. In epidemiology, for example, the expression $P(Y_{x'} = y'|x, y)$ may stand for the fraction of patients who recovered (y) under treatment (x) that would not have recovered (y') had they not been treated (x'). This fraction cannot be assessed in experimental study, for the simple reason that we cannot re-test patients twice, with and without treatment. A different question is therefore posed: which counterfactuals can be tested, be it in experimental or observa-

tional studies. This question has been given a mathematical solution in (Shpitser and Pearl, 2007). It has been shown, for example, that in linear systems, $E(Y_x|e)$ is estimable from experimental studies whenever the prospective effect $E(Y_x)$ is estimable in such studies.

Retrospective counterfactuals have also been indispensable in conceptualizing direct and indirect effects (Baron and Kenny, 1986; Robins and Greenland, 1992; Pearl, 2001), which require nested counterfactuals in their definitions. For example, to evaluate the direct effect of treatment $X = x'$ on individual u, un-mediated by a set Z of intermediate variables, we need to construct the nested counterfactual $Y_{x',Z_x(u)}$ where Y is the effect of interest, and $Z_x(u)$ stands for whatever values the intermediate variables Z would take had treatment not been given. Likewise, the average *indirect effect*, of a transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z, hypothetically, to whatever value it would have attained had X been set to $X = x'$.

This formalism has enabled researchers to derive conditions under which direct and indirect effects are estimable from empirical data (Pearl, 2001; Petersen et al., 2006) and to answer such questions as: "Can data prove an employer guilty of hiring discrimination?" or, using the classical example of Hesslow (1976) and Cartwright (1989) "Can data help determine the direct effect of a birth-control pill on thrombosis, unmediated by pregnancy?"[4]

The impact of the structural theory in the empirical sciences does not prove, of course, its merits as a cognitive theory of counterfactual reasoning. It proves nevertheless that in the arena of policy evaluation and decision making the theory is compatible with investigators

[4]Note that conditioning on the intermediate variables in Z would generally yield the wrong answer, due to unobserved "confounders" affecting both Z and Y. Moreover, in non linear systems, the value at which we hold Z constant will affect the result (Pearl, 2000a, pp. 126-132).

states of belief and, whenever testable, its conclusions have withstood the test of fire.

4.3 from kyono (r364) – need to format:fig1ab, sec2-r364 (not in latex)

2 Overview

This section presents a summary of the six tasks that Commentator performs.

(a) (b)

Figure 1: (a) A Sample Path Diagram [Pearl, 2009]. (Note: latent variables are not shown

2.1 Task 1: Find All Minimal Testable Implications

Input: A DAG G and a list of observable and latent variables.

Output: A list of the minimal testable implications of the model, specifically, a list o

Example: Consider Figure 1 as input, Commentator produces the following list:

? ??????1=0

? ??????2=0

? ????????1??1=0

? ????????1??2??1??2=0

? ????1??2??2=0

V

W2

W1

Z1

X

Y

Z2

B

A

C

e1

e 2

e 3

e 4

e 7

e 5

e 6

4

? ????2??1??1=0

? ??????2????2=0

? ????1??2????1=0

? ????1??2????2=0

? ????1??1????1??2=????1??2

? ????2??2????1??2=????2??1

? ????2??2????1??=????2??

? ????1????1????1??=??????

? ????2??????2????2????=????????

2.2 Task 2: Find Identifiable Path Coefficients

Input: A DAG G and a list of observable and latent variables.

Output: List of all path coefficients that are identifiable using simple regression.

Example: Consider Figure 1 as input, Commentator produces the following list:

? The coefficient ? on V?Z1 is identifiable controlling for the Empty Set, that is ?=???

? The coefficient ? on V?Z2 is identifiable controlling for the Empty Set, that is ?=???

? The coefficient ? on W1?Z1 is identifiable controlling for the Empty Set, that is ?=??

? The coefficient ? on W2?Z2 is identifiable controlling for the Empty Set, that is ?=??

? The coefficient ? on Z1?X is identifiable controlling for W1, that is ?=????1????1

? The coefficient ? on Z2?Y is identifiable controlling for V W2, that is ?=????2??????2

2.3 Task 3: Find Identifiable Total Effects

Input: A DAG G and a list of observable and latent variables.

Output: List of total effects that are identifiable using simple regression.

Example: Consider Figure 1 as input, Commentator produces the following list:

? The total effect t of V on X is identifiable controlling for the Empty Set, that is ??

? The total effect t of V on Y is identifiable controlling for the Empty Set, that is ??

? The total effect t of V on Z1 is identifiable controlling for the Empty Set, that is ?

5

? The total effect t of V on Z2 is identifiable controlling for the Empty Set, that is ?

? The total effect t of W1 on Z1 is identifiable controlling for the Empty Set, that is

? The total effect t of W2 on Z2 is identifiable controlling for the Empty Set, that is

? The total effect t of Z1 on X is identifiable controlling for W1, that is ??=????1????

? The total effect t of Z1 on Y is identifiable controlling for V W1, that is ??=????1??

? The total effect t of Z2 on Y is identifiable controlling for V W2, that is ??=?????2??

2.4 Task 4: Find All Instrumental Variables

Input: A DAG G and a list of observable and latent variables.

Output: List of all instrumental variables and the parameters they help identify.

Example: Consider Figure 1 as input, the Commentator produces the following list:

? The coefficient ? on W1?Z1 is identifiable via instrumental variable W2 controlling fo

? The coefficient ? on W2?Z2 is identifiable via instrumental variable W1 controlling fo

? The coefficient ? on W2?Z2 is identifiable via instrumental variable X controlling for

? The coefficient ? on X?Y is identifiable via instrumental variable Z1 controlling for

? The coefficient ? on Z1?X is identifiable via instrumental variable V controlling for

? The coefficient ? on Z2?Y is identifiable via instrumental variable V controlling for

2.5 Task 5: Test for Model Equivalence and Nestedness

Input: Two DAGs, G and G, and a list of observable and latent variables.

Output: Determine whether G is equivalent to G, G is nested in G, or vice versa. (This i

Figure 2: A Sample Path Diagram [Pearl, 2009].

(Note: latent variables are not shown explicitly,

but are presumed to emit the double-head arrows.)

Example: Consider Figure 1 and 2 as input, Commentator produces the output shown in Tabl

Table 1: Commentator Output: Testable Implications for Figure 1 and 2

Commentator Output for Figure 2

? ??????1=0

? ??????2=0

? ????????1??1=0

? ????????1??2??1??2=0

? ????????1??2????2=0

? ????1??2??2=0

? ????2??1??1=0

? ??????2????2=0

? ????1??2????1=0

? ????1??2????2=0

? ??????1????1??2??=0

? ??????1??1??2????2= 0

Commentator Output for Figure 1

? ??????1=0

? ??????2=0

? ????????1??1=0

? ????????1??2??1??2=0

? ????1??2??2=0

? ????2??1??1=0

? ??????2????2=0

? ????1??2????1=0

? ????1??2????2=0

2.6 Task 6: Find all Minimum-Size Admissible Sets of Covariates for Determining the Caus

Input: A DAG G, a list of observable and latent variables, and a pair of variables, X an

Output: List of all minimum-size admissible sets S, such that conditioning on S removes

Example 1: Consider Figure 17(b) as input, Commentator produces the following output:

? The causal effect ??(??|????(??)) can be estimated by adjustment on:

o V, W1, W2

o W1, W2, Z1

o W1, W2, Z2

Example 2: Consider Figure 17(a) as input, Commentator produces the following output:

? The causal effect ??(??|????(??)) cannot be identified by adjustment, i.e., no admissi

sec3.1, 4.3.1, sec6-r355

4.4 Introduction to Structural Equation Models (sec3.1-r355)

How can we express mathematically the common understanding that symptoms do not cause diseases? The earliest attempt to formulate such relationship mathematically was made in the 1920s by the geneticist Sewall Wright (1921). Wright used a combination of equations and graphs to communicate causal relationships. For example, if X stands for a disease variable and Y stands for a certain symptom of the disease, Wright would write a linear equation:

$$y = \beta x + u_Y, \tag{3}$$

where x stands for the level (or severity) of the disease, y stands for the level (or severity) of the symptom, and u_Y stands for all factors, other than the disease in question, that could possibly affect Y when X is held constant.[5] In interpreting this equation we should think of a physical process whereby nature *examines* the values of x and u and, accordingly, *assigns*

[5]Linear relations are used here for illustration purposes only; they do not represent typical disease-symptom relations but illustrate the historical development of path analysis. Additionally, we will use standardized variables—that is, zero mean and unit variance.

to variable Y the value $y = \beta x + u_Y$. Similarly, to "explain" the occurrence of disease X, we could write $x = u_X$, where U_X stands for all factors affecting X.

Equation (3) still does not properly express the causal relationship implied by this assignment process, because algebraic equations are symmetrical objects; if we rewrite (3) as

$$x = (y - u_Y)/\beta, \tag{4}$$

it might be misinterpreted to mean that the symptom influences the disease. To express the directionality of the underlying process, Wright augmented the equation with a diagram, later called "path diagram," in which arrows are drawn from (perceived) causes to their (perceived) effects, and more importantly, the absence of an arrow makes the empirical claim that Nature assigns values to one variable irrespective of another. In Figure 3, for example, the absence of arrow from Y to X represents the claim that symptom Y is not among the factors U_X that affect disease X. Thus, in our example, the complete model of a symptom and a disease would be written as in Figure 3: 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 parameter β in the equation is called a "path coefficient," and it 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, and regardless of whether the increase in X originates from external or internal influences.

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 are sometimes called "disturbances" or "errors"; they represent factors omitted from the model but judged to be relevant for explaining the behavior of variables in the model. Variable U_X, for example, represents factors that contribute to the disease X, which may or may not be correlated with U_Y (the factors that influence the symptom Y). Thus, background factors in structural equations 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 part of physical reality (e.g., genetic factors, socioeconomic conditions), which are responsible for variations observed in the data; 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 3(b). By allowing correlations among omitted factors, we allow in effect for the presence of latent variables affecting both X and Y, as shown explicitly in Figure 3(c), which is the standard representation in the SEM literature (e.g., Bollen, 1989). In contrast to traditional latent variable models, however, our attention will not be focused on the connections among such latent variables but, rather, on the causal effects that those variables induce among the observed variables. In particular, we will not be interested in the causal effect of one latent variable on another and, therefore, there will be no reason to distinguish between correlated errors and causally related latent variables;

it is only the distinction between correlated and uncorrelated errors (e.g., between Figure 3(a) and (b)) that need to be made. Moreover, when the error terms are uncorrelated, it is often more convenient to eliminate them altogether from the diagram (as in Figure **??**, Section **??**), with the understanding that every variable, X, is subject to the influence of an independent disturbance U_X.

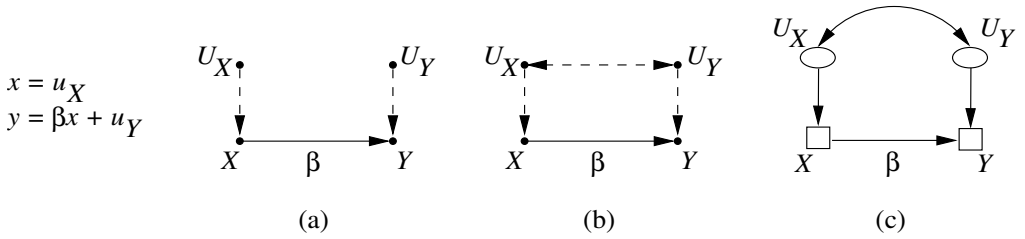

Figure 3: A simple structural equation model, and its associated diagrams, showing (a) independent unobserved exogenous variables (connected by dashed arrows), (b) dependent exogenous variables, and (c) an equivalent, more traditional notation, in which latent variables are enclosed in ovals.

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 3(b), one consisting of the direct arrow $X \rightarrow Y$ while the other tracing the nodes X, U_X, U_Y, and Y.

Wright's major contribution to causal analysis, aside from introducing the language of path diagrams, has been the development of graphical rules for writing down the covariance of any pair of observed variables in terms of path coefficients and of covariances among the

error terms. In our simple example, we can immediately write the relations

$$Cov(X, Y) = \beta \tag{5}$$

for Figure 3(a), and

$$Cov(X, Y) = \beta + Cov(U_Y, U_X) \tag{6}$$

for Figure 3(b)–(c). (These can be derived of course from the equations, but, for large models, algebraic methods tend to obscure the origin of the derived quantities). Under certain conditions, (e.g., if $Cov(U_Y, U_X) = 0$),such relationships may allow us to solve for the path coefficients in terms of observed covariance terms only, and this amounts to inferring the magnitude of (direct) causal effects from observed, nonexperimental associations, assuming of course that we are prepared to defend the causal assumptions encoded in the diagram.

It is important to note that, 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. In Figure 3(a), for example, the assumptions that permit us to identify the direct effect β are encoded by the missing double arrow between U_X and U_Y, indicating $Cov(U_Y, U_X)$=0, together with the missing arrow from Y to X. Had any of these two links been added to the diagram, we would not have been able to identify the direct effect β. Such additions would amount to relaxing the assumption $Cov(U_Y, U_X) = 0$, or the assumption that Y does not effect X, respectively. Note also that both assumptions are causal, not statistical, since none can be determined from the joint density of the observed variables, X and Y; the association between the unobserved terms, U_Y and U_X, can only be uncovered

in an experimental setting; or (in more intricate models, as in Figure **??**) from other causal assumptions.

Although each causal assumption in isolation cannot be tested, the sum total of all causal assumptions in a model often has testable implications. The chain model of Figure 5(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 those 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 1 (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 X to Y, it is said to "d-separate X and Y," and then, X and Y are independent given S, written $X \perp\!\!\!\perp Y|S$.*

To illustrate, the path $U_Z \rightarrow Z \rightarrow X \rightarrow Y$ 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 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.

5 section 5

5.1 Equivalent Models

d-separation also defines conditions for model equivalence that are easily ascertained in DAGs (Verma and Pearl, 1990) and DAGs with latent variables (Ali et al., 2009). These conditions are mathematically proven and should therefore supercede the heuristic (and occasionally faulty) rules prevailing in SEM's research (Lee and Hershberger, 1990).

For example, Consider the following model

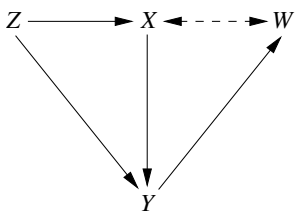

Figure 4: Showing discrepency between Lee and Hershberger's replacement rule and d-separation, which forbid the replacement of $X \rightarrow Y$ by $X \leftrightarrow Y$.

According to the replacement criterion of Lee and Hershberger we can replace the arrow $X \rightarrow Y$ with a double arrow edge $X \leftrightarrow Y$. (representing residual correlation) The reason being that the predictors (Z) of the effect variable (Y) are the same as those for the source variable (X) Unfortunately, the post-replacement model imposes additional constraint $b_{WZ \cdot Y} = 0$, that is not imposed by the pre-replacement model . This can be seen

from the fact that the path $Z - -Y - -X - -W$ is blocked by Y in the post-replacement model and not in the original model.

The conditional independencies entailed by d-separation constitute the main opening through which the assumptions embodied in structural equation models can confront the scrutiny of nonexperimental data. In other words, almost all statistical tests capable of invalidating the model are entailed by those implications.[6] In addition, d-separation further

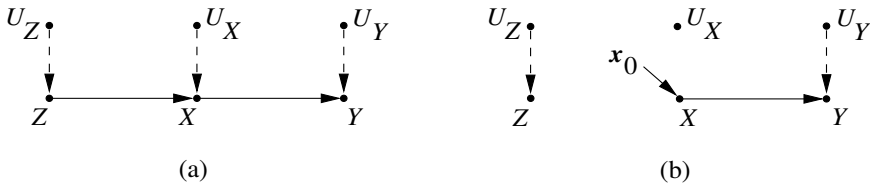

Figure 5: The diagrams associated with (a) the structural model of equation (**??**) and (b) the modified model of equation (**??**), representing the intervention $do(X = x_0)$.

defines conditions for model equivalence (Verma and Pearl, 1990; Ali et al., 2009) that are mathematically proven and should therefore supercede the heuristic (and faulty) rules prevailing in social science research (Lee and Hershberger, 1990).

5.1.1 Testing the Relevant Assumptions (sec4.3.1-r355)

When Q is identifiable, the structural framework also delivers an algebraic expression for the estimand $EST(Q)$ of the target quantity Q, examples of which are given in equations (**??**) and (**??**), and estimation techniques are then unleashed as discussed in Section **??**. A prerequisite part of this estimation phase is a test for the testable implications, if any, of those assumptions in M that render Q identifiable—there is no point in estimating $EST(Q)$

[6] Additional implications called "dormant independence" (Shpitser and Pearl, 2008) may be deduced from some semi-Markovian models, i.e., graphs with correlated errors (Verma and Pearl, 1990).

if the data proves those assumptions false and $EST(Q)$ turns out to be a misrepresentation of Q. 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 tested in the data to support the validity of M. These can easily be enumerated by attending to each missing edge in the graph. For example, in Figure **??**, the missing edges are $Z_1 - Z_2, Z_1 - Y$, and $Z_2 - X$. Accordingly, the testable implications of M are

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

In linear systems, these conditional independence constraints translate into zero coefficients in the proper 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, 2009b, pp. 143–45). Software 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

unobserved confounders, then additional constraints can be tested, beyond the d-separation conditions (see footnote 6).

Investigators should be reminded, however, that only a fraction, called "kernel," of the assumptions embodied in M are needed for identifying Q (Pearl, 2004), the rest may be violated in the data with no effect on Q. In Figure **??**, for example, the assumption $\{U_Z \perp\!\!\!\perp U_X\}$ is not necessary for identifying $Q = P(y|do(x))$; the kernel $\{U_Y \perp\!\!\!\perp U_Z, U_Y \perp\!\!\!\perp U_X\}$ (together with the missing arrows) is sufficient. Therefore, the testable implication of this kernel, $Z \perp\!\!\!\perp Y|X$, is all we need to test when our target quantity is Q; the assumption $\{U_Z \perp\!\!\!\perp U_X\}$ need not concern us.

More importantly, investigators must keep in mind that only a tiny fraction of any kernel lends itself to statistical tests; the bulk of it must remain untestable, at the mercy of scientific judgment. In Figure **??**, for example, the assumption set $\{U_X \perp\!\!\!\perp U_Z, U_Y \perp\!\!\!\perp U_X\}$ constitutes a sufficient kernel for $Q = P(y|do(x))$ (see equation **??**) yet it has no testable implications whatsoever. The prevailing practice of submitting an entire structural equation model to a "goodness of fit" test (Bollen, 1989) in support of causal claims is at odds with the logic of SCM (see Pearl, 2000a, pp. 144–45. Statistical tests can be used for rejecting certain kernels in the rare cases where such kernels have testable implications, but passing these tests does not prove the validity of any causal claim; one can always find alternative causal models that make a contradictory claim and, yet, possess identical statistical implications.[7] The lion's

[7]This follows logically from the demarcation line of Section **??**. The fact that some social scientists were surprised by the discovery of contradictory equivalent models (see (Pearl, 2009b, p. 148) suggests that these scientists did not take very seriously the ramifications of the causal-statistical distinction, or that they misunderstood the conditional nature of all causal claims drawn from observational studies (see Pearl, 2009b,

share of supporting causal claims falls on the shoulders of untested causal assumptions.[8]

Some researchers consider this burden to be a weakness of SCM and 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 (2006) imply that SCM operates at the boundary of this universal relationship; no method can do better.

5.2 Mediation: Direct and Indirect Effects (sec6-r355)

5.2.1 Direct Versus Total Effects

The causal effect we have analyzed so far, $P(y|do(x))$, measures the *total* effect of a variable (or a set of variables) X on a response variable Y. In many cases, this quantity does not adequately represent the target of investigation and attention is focused instead on the *direct* effect of X on Y. The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$,

pp. 369–73.

[8]The methodology of "causal discovery" (Spirtes et al. 2000; Pearl 2000a, ch. 2) is likewise based on the causal assumption of "faithfulness" or "stability"—a problem-independent assumption that constrains the relationship between the structure of a model and the data it may generate. We will not assume stability in this paper.

which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

From a policymaking viewpoint, an investigator may be interested in decomposing effects to quantify the extent to which racial salary disparity is due to educational disparity, or, more generally, the extent to which sensitivity to a given variable can be reduced by eliminating sensitivity to an intermediate factor, standing between that variable and the outcome. Often, 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.

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 (Alwin and Hauser, 1975; Graff and Schmidt, 1981; Sobel, 1987; 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 (Judd and Kenny, 1981; Baron and Kenny, 1986; Muller et al., 2005; Shrout and Bolger, 2002; MacKinnon et al., 2007a) primarily because structural equation modeling in those sciences were deeply entrenched in linear analysis, where the distinction

between causal parameters and their regressional interpretations can easily be conflated. 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., 2007b). These difficulties have accentuated the need to redefine and derive causal effects from first principles, uncommitted to distributional assumptions or a particular parametric form of the equations. The structural methodology presented in this paper adheres to this philosophy and it has produced indeed a principled solution to the mediation problem, based on the counterfactual reading of structural equations (1). The subsections, that follow summarize the method and its solution.

5.2.2 Controlled Direct Effects

A major impediment to progress in mediation analysis has been the lack of notational facility for expressing the key notion of "holding the mediating variables fixed" in the definition of direct effect. Clearly, this notion must be interpreted as (hypothetically) setting the intermediate variables to constants by physical intervention, not by analytical means such as selection, regression conditioning, matching, or adjustment. For example, consider the

simple mediation models of Figure 6(a), which reads

$$\begin{aligned} x &= u_X \\ Z &= f_Z(x, u_Z) \\ y &= f_Y(x, z, u_Y) \end{aligned} \tag{7}$$

and where the error terms (not shown explicitly) are assumed to be mutually independent.

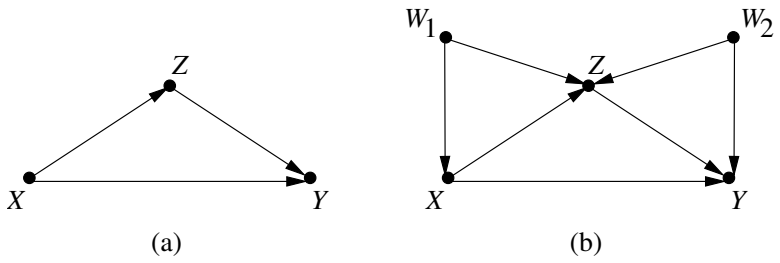

Figure 6: A generic model depicting mediation through Z (a) with no confounders and (b) with two confounders, W_1 and W_2.

To measure the direct effect of X on Y it is sufficient to measure their association conditioned on the mediator Z. In Figure 6(b), however, where the error terms are dependent, it will not be sufficient to measure the association between X and Y for a given level of Z because, by conditioning on the mediator Z, which is a collision node (Definition 1), we create spurious associations between X and Y through W_2, even when there is no direct effect of X on Y (Pearl, 1998; Cole and Hernán, 2002).[9]

Using the $do(x)$ notation, enables us to correctly express the notion of "holding Z fixed" and to obtain a simple definition of the *controlled direct effect* of the transition from $X = x$

[9]The need to control for mediator-outcome confounders (e.g., W_2 in Figure 6(b)) was evidently overlooked in the classical paper of Baron and Kenny (1986), and has subsequently been ignored by most social science researchers.

to $X = x'$:

$$CDE \triangleq E(Y|do(x'), do(z)) - E(Y|do(x), do(z)).$$

Or, equivalently, we can use counterfactual notation

$$CDE \triangleq E(Y_{x'z}) - E(Y_{xz}),$$

where Z is the set of all mediating variables. Readers can easily verify that, in linear systems, the controlled direct effect reduces to the path coefficient of the link $X \rightarrow Y$ (see footnote **??**) regardless of whether confounders are present (as in Figure 6(b)) and regardless of whether the error terms are correlated or not.

This separates the task of definition from that of identification, as demanded by Section **??**. The identification of CDE would depend, of course, on whether confounders are present and whether they can be neutralized by adjustment, but these do not alter its definition. Nor should trepidation about infeasibility of the action $do(gender = male)$ enter the definitional phase of the study. Definitions apply to symbolic models, not to human biology.[10]

Graphical identification conditions for multi-action expressions of the type $E(Y|do(x), do(z_1), do(z_2), \ldots, do(z_k))$ in the presence of unmeasured confounders were derived by Pearl and Robins (1995) (see Pearl, 2000a, ch. 4) using sequential application of the back-door conditions discussed in Section **??**.

[10]In reality, it is the employer's perception of applicant's gender and his or her assessment of gender-job compatibility that renders gender a "cause" of hiring; manipulation of gender is not needed.

5.3 Natural Direct Effects

In linear systems, the direct effect is fully specified by the path coefficient attached to the link from X to Y; therefore, the direct effect is independent of the values at which we hold Z. In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. For example, it is not uncommon to find employers who prefer males for the high-paying jobs (i.e., high z) and females for low-paying jobs (low z).

When the direct effect is sensitive to the levels at which we hold Z, it is often more meaningful to define the direct effect relative to some "natural" base-line level that may vary from individual to individual, and represents the level of Z just before the change in X. Conceptually, we can define the natural 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 $do(x)$. This hypothetical change, which Robins and Greenland (1992) conceived and called "pure" and Pearl (2001) formalized and analyzed under the rubric "natural," mirrors what lawmakers instruct us to consider in race or sex discrimination cases: "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (In *Carson versus Bethlehem Steel Corp.*, 70 FEP Cases 921, 7th Cir. (1996)). Thus, whereas the controlled direct effect measures the effect of X on Y while holding Z fixed at a uniform

[11] Pearl (2001) used the acronym NDE to denote the natural direct effect. We will delete the letter "N" from the acronyms of both the direct and indirect effect, and use DE and IE, respectively.

level (z) for all units,[12] the natural direct effect allows z to vary from individual to individual to be held fixed at whatever level each individual obtains naturally, just before the change in X.

Extending the subscript notation to express nested counterfactuals, Pearl (2001) gave the following definition for the "natural direct effect":

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \tag{8}$$

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 $X = x$. We see that $DE_{x,x'}(Y)$, the natural direct effect of the transition from x to x', involves probabilities of *nested counterfactuals* and cannot be written in terms of the $do(x)$ operator. Therefore, the natural direct effect cannot in general be identified or estimated, even with the help of ideal, controlled experiments (see footnote **??** for intuitive explanation). However, aided by the surgical definition of equation (1) and the notational power of nested counterfactuals, Pearl (2001) was nevertheless able to show that, if certain assumptions of "no confounding" are deemed valid, the natural direct effect can be reduced to

$$DE_{x,x'}(Y) = \sum_z [E(Y|do(x', z)) - E(Y|do(x, z))]P(z|do(x)). \tag{9}$$

The intuition is simple; the natural direct effect is the weighted average of the controlled direct effect, using the causal effect $P(z|do(x))$ as a weighing function.

One condition for the validity of (9) is that $Z_x \perp\!\!\!\perp Y_{x',z}|W$ holds for some set W of measured covariates. This technical condition in itself, like the ignorability condition of (**??**), is

[12] In the hiring discrimination example, this would amount, for example, to testing gender bias by marking all application forms with the same level of schooling and other skill-defining attributes.

close to meaningless for most investigators, as it is not phrased in terms of realized variables. The surgical interpretation of counterfactuals (1) can be invoked at this point to unveil the graphical interpretation of this condition. It states that W should be admissible (i.e., satisfy the back-door condition) relative to the path(s) from Z to Y. This condition, satisfied by W_2 in Figure 6(b), is readily comprehended by empirical researchers, and the task of selecting such measurements, W, can then be guided by available scientific knowledge. Additional graphical and counterfactual conditions for identification are derived in Pearl (2001), Petersen et al. (2006), and Imai et al. (2008).

In particular, it can be shown (Pearl, 2001) that expression (9) 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 (**??**) or (**??**) and then estimated by regression.

For example, for the model in Figure 6(b), equation (9) reads

$$DE_{x,x'}(Y) = \sum_z \sum_{w_2} P(w_2)[E(Y|x', z, w_2)) - E(Y|x, z, w_2))] \sum_{w_1} P(z|x, w_1)P(w_1). \quad (10)$$

while for the confounding-free model of Figure 6(a) we have

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

Both (10) and (11) can easily be estimated by a two-step regression.

5.4 Natural Indirect Effects

Remarkably, the definition of the natural direct effect (8) can be turned around and provide an operational definition for the *indirect effect*—a concept shrouded in mystery and con-

troversy, because it is impossible, using any physical intervention, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

The *natural 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

$$IE_{x,x'}(Y) \triangleq E[(Y_{x,Z_{x'}}) - E(Y_x)], \tag{12}$$

which is almost identical to the direct effect (equation 8) 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,

$$TE_{x,x'}(Y) \triangleq E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{13}$$

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

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{14}$$

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.[13]

Note that, although it cannot be expressed in *do*-notation, the indirect effect has clear policymaking implications. For example, in the hiring discrimination context, a policymaker

[13]Some authors (e.g., VanderWeele, 2009), define the natural indirect effect as the difference $TE - DE$. This renders the additive formula a tautology of definition, rather then a theorem, conditioned upon the anti-symmetry $IE_{x,x'}(Y) = -IE_{x',x}(Y)$. Violation of (14) will be demonstrated in the next section.

may be interested in predicting the gender mix in the workforce if gender bias is eliminated and all applicants are treated equally—say, the same way that males are currently treated. This quantity will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender-dependent.

More generally, a policymaker may be interested in the effect of issuing a directive to a select set of subordinate employees, or in carefully controlling the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*—that is, the effect of X on Y through a selected set of paths (Avin et al., 2005).

In all these cases, the policy intervention invokes the selection of signals to be sensed rather than variables to be fixed. Pearl (2001) has therefore suggested that *signal sensing* is more fundamental to the notion of causation than *manipulation*; the latter being but a crude way of stimulating the former in an experimental setup. The mantra "No causation without manipulation" must be rejected (see Pearl, 2009b, sec. 11.4.5).

It is remarkable that counterfactual quantities like DE and IE, which could not be expressed in terms of $do(x)$ operators and therefore appear void of empirical content, can under certain conditions be estimated from empirical studies, and serve to guide policies. Awareness of this potential should embolden researchers to go through the definitional step of the study and freely articulate the target quantity $Q(M)$ in the language of science—that is, structure-based counterfactuals—despite the seemingly speculative nature of each assumption in the model (Pearl, 2000b).

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

This subsection demonstrates how the solution provided in equations (11) and (14) can be applied in assessing mediation effects in nonlinear models. We will use the simple mediation model of Figure 6(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 10) and that integrals should replace summations when dealing with continuous variables (Imai et al., 2008).

Combining (9) and (14), the expression for the indirect effect, IE, becomes

$$IE_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)] \tag{15}$$

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 will refer to this expression as the "Mediation Formula."

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 6(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 coun-

terfactual 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

$$g(x, z) = E(Y|x, z)$$

for every (x, z) cell. In the second step we estimate the conditional expectation of $g(x, z)$ with respect to z, conditional on $X = x'$ and $X = x$, respectively, and take the difference

$$IE_{x,x'}(Y) = E_z(g(x, z)|x') - E_z(g(x, z)|x).$$

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 (15) 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 6(a) (equation 7), which reads

$$\begin{aligned} x &= u_X \\ z &= b_0 + b_x x + u_Z \\ y &= c_0 + c_x x + c_z z + u_Y \end{aligned} \tag{16}$$

with u_X, u_Y, and u_Z uncorrelated, zero-mean error terms. Computing the conditional ex-

pectation in (15) gives

$$E(Y|x,z) = E(c_0 + c_x x + c_z z + u_Y) = c_0 + c_x x + c_z z$$

and yields

$$\begin{aligned} IE_{x,x'}(Y) &= \sum_z (c_x x + c_z z)[P(z|x') - P(z|x)]. \\ &= c_z[E(Z|x') - E(Z|x)] \qquad (17) \\ &= (x' - x)(c_z b_x) \qquad (18) \\ &= (x' - x)(b - c_x) \qquad (19) \end{aligned}$$

where b is the total effect coefficient,

$$b = (E(Y|x') - E(Y|x))/(x' - x) = c_x + c_z b_x.$$

We thus obtained the standard expressions for indirect effects in linear systems, which can be estimated either as a difference in two regression coefficients (equation 19) or a product of two regression coefficients (equation 18), with Y regressed on both X and Z (see MacKinnon et al., 2007b). These two strategies do not generalize to nonlinear systems as shown in Pearl (2010a); direct application of (15) is necessary.

To understand the difficulty, consider adding an interaction term $c_{xz}xz$ to the model in equation (16), yielding

$$y = c_0 + c_x x + c_z z + c_{xz} xz + u_Y$$

Now assume that, through elaborate 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 $c_z b_x/b$ (equation 18), the latter by the difference $(b - c_x)/b$ (equation 19) 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 (11), (14) and (15) and assuming $x = 0$ and $x' = 1$, yields the following decomposition:

$$\begin{aligned} DE &= c_x + b_0 c_{xz} \\ IE &= b_x c_z \\ TE &= c_x + b_0 c_{xz} + b_x(c_z + c_{xz}) \\ &= DE + IE + b_x c_{xz} \end{aligned}$$

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

$$IE/TE = b_x c_z/(c_x + b_0 c_{xz} + b_x(c_z + c_{xz}))$$

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

$$1 - DE/TE = b_x(c_z + c_{xz})/(c_x + b_0 c_{xz} + b_x(c_z + c_{xz}))$$

We note that, due to interaction, a direct effect can be sustained even when the parameter c_x 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 effect 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 impact 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 multivalued outcomes are straightforward.

Assume that the model of Figure 6(a) is valid and that the observed data is given by Figure 7. The factors $E(Y|x,z)$ and $P(Z|x)$ can be readily estimated as shown in the two right-most columns of Figure 7 and, when substituted in (11), (14), (15), yield

$$DE = (g_{10} - g_{00})(1 - h_0) + (g_{11} - g_{01})h_0 \tag{20}$$

$$IE = (h_1 - h_0)(g_{01} - g_{00}) \tag{21}$$

$$TE = g_{11}h_1 + g_{10}(1 - h_1) - [g_{01}h_0 + g_{00}(1 - h_0)] \tag{22}$$

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

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		

Figure 7: Computing the Mediation Formula for the model in Figure 6(a), with X, Y, Z binary.

process.

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 equation (20).

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 (15) explicates this target quantity formally, and casts it in terms of estimable quantities.

The derivation of the Mediation Formula was facilitated by taking seriously the five steps of the structural methodology (Section **??**) together with the graphical-counterfactual-structural symbiosis spawned by the surgical interpretation of counterfactuals (equation 1).

In contrast, when the mediation problem is approached from an exclusivist potential-outcome viewpoint, void of the structural guidance of equation (1), 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 (1).[14]

6 Structural Models, Diagrams, Causal Effects, and Counterfactuals

This section provides a gentle introduction to the structural framework and uses it to present the main advances in causal inference that have emerged in the past two decades. We start with recursive linear models,[15] in the style of Wright (1923), Blalock (1964), and Duncan (1975) and, after explicating carefully the meaning of each symbol and the causal assumptions embedded in each equation, we advance to nonlinear and nonparametric models with latent variables, and we show how these models facilitate a general analysis of causal effects and counterfactuals.

References

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

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

[15]By "recursive" we mean systems free of feedback loops. We allow however correlated errors, or latent variables that create such correlations. Most of our results, with the exception of Sections **??** and **??** are valid for nonrecursive systems, allowing reciprocal causation.

Alwin, D. and Hauser, R. (1975). The decomposition of effects in path analysis. *American Sociological Review* **40** 37–47.

Avin, C., Shpitser, I. and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*. Morgan-Kaufmann Publishers, Edinburgh, UK.

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.

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

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford.

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.

Cole, S. and Hernán, M. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31** 163–165.

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.

Gardenfors, P. (1988). Causation and the dynamics of belief. In *Causation in Decision, Belief Change and Statistics II* (W. Harper and B. Skyrms, eds.). Kluwer Academic Publishers, 85–104.

Graff, J. and Schmidt, P. (1981). A general model for decomposition of effects. In *Systems Under Indirect Observation, Part 1* (K. Jöreskog and H. Wold, eds.). North-Holland, Amsterdam, 131–148.

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.

Hesslow, G. (1976). Discussion: Two notes on the probabilistic approach to causality. *Philosophy of Science* **43** 290–292.

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

Hurwicz, L. (1950). Generalization of the concept of identification. In *Statistical Inference in Dynamic Economic Models* (T. Koopmans, ed.). Cowles Commission, Monograph 10, Wiley, New York, 245–257.

Imai, K., Keele, L. and Yamamoto, T. (2008). Identification, inference, and sensitivity analysis for causal mediation effects. Tech. rep., Department of Politics, Princton University. Forthcoming *Statistical Science*.

Joyce, J. (1999). *The Foundatins of Causal Decision Theory*. Cambridge University Press, Cambridge, MA.

Judd, C. and Kenny, D. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* **5** 602–619.

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.

Lewis, D. (1973). Counterfactuals and comparative possibility. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.). *Ifs*, D. Reidel, Dordrecht, pages 57–85, 1981.

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

MacKinnon, D., Fairchild, A. and Fritz, M. (2007a). Mediation analysis. *Annual Review of Psychology* **58** 593–614.

MacKinnon, D., Lockwood, C., Brown, C., Wang, W. and Hoffman, J. (2007b). 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.

Marschak, J. (1953). Studies in econometric method. In *Economic Measurements for Policy and Prediction* (W. C. Hood and T. Koopmans, eds.). Cowles Commission Monograph 10, Wiley and Sons, Inc., 1–26.

Muller, D., Judd, C. and Yzerbyt, V. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology* **89** 852–863.

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

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

Pearl, J. (2000b). Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association* **95** 428–431.

Pearl, J. (2001). Direct and indirect effects. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*. 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. (2009a). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146, <http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>.

Pearl, J. (2009b). *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.

Pearl, J. and Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, 444–453.

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.

Shpitser, I. and Pearl, J. (2006). 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. (2007). What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Vancouver, BC, Canada, 352–359. Also, *Journal of Machine Learning Research*, 9:1941–1979, 2008.

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.

Shrout, P. and Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods* **7** 422–445.

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.

Skyrms, B. (1980). *Causal Necessity*. Yale University Press, New Haven.

Sobel, M. (1987). Direct and indirect effects in linear structural equation models. *Sociological Methods & Research* **16** 1155–176.

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.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search*. 2nd ed. MIT Press, Cambridge, MA.

Stalnaker, R. (1972). Letter to David Lewis. In W.L. Harper, R. Stalnaker, and G. Pearce (Eds.), *Ifs*, D. Reidel, Dordrecht, pages 151–152, 1981.

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*. 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.

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, 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.