Interpretation and Identification of Causal Mediation

Judea Pearl
University of California, Los Angeles

This article reviews the foundations of causal mediation analysis and offers a general and transparent account of the conditions necessary for the identification of natural direct and indirect effects, thus facilitating a more informed judgment of the plausibility of these conditions in specific applications. I show that the conditions usually cited in the literature are overly restrictive and can be relaxed substantially without compromising identification. In particular, I show that natural effects can be identified by methods that go beyond standard adjustment for confounders, applicable to observational studies in which treatment assignment remains confounded with the mediator or with the outcome. These identification conditions can be validated algorithmically from the diagrammatic description of one’s model and are guaranteed to produce unbiased results whenever the description is correct. The identification conditions can be further relaxed in parametric models, possibly including interactions, and permit one to compare the relative importance of several pathways, mediated by interdependent variables.

**Keywords:** mediation formula, identification, confounding, graphical models

Mediation analysis aims to uncover causal pathways along which changes are transmitted from causes to effects. Interest in mediation analysis stems from both scientific and practical considerations. Scientifically, mediation tells us how nature works, and practically, it enables us to predict behavior under a rich variety of conditions and policy interventions. For example, in coping with the age-old problem of gender discrimination (Bickel, Hammel, & O’Connell, 1975; Goldberger, 1984), a policymaker may be interested in assessing the extent to which gender disparity in hiring can be reduced by making hiring decisions gender-blind, compared with eliminating gender inequality in education or job qualifications. The former concerns the direct effect of gender on hiring, while the latter concerns the indirect effect or the effect mediated via job qualification.

The example illustrates two essential ingredients of modern mediation analysis. First, the indirect effect is not merely a modeling artifact formed by suggestive combinations of parameters but an intrinsic property of reality that has tangible policy implications. In this example, reducing employers’ prejudices and launching educational reforms are two contending policy options that involve costly investments and different implementation efforts. Knowing in advance which of the two, if successful, has a greater impact on reducing hiring disparity is essential for planning and depends critically on mediation analysis for resolution. Second, the policy decisions in this example concern the enabling and disabling of processes (hiring vs. education) rather than lowering or raising values of specific variables. These two considerations lead to the analysis of natural direct and indirect effects.

Mediation analysis has its roots in the literature of structural equation models (SEMs), going back to Wright’s (1923, 1934) method of path analysis and continuing in the social sciences from the 1960s to 1980s through the works of Baron and Kenny (1986), Bollen (1989), Duncan (1975), and Fox (1980). The bulk of this work was carried out in the context of linear models, in which effect sizes are represented as sums and products of structural coefficients. The definition, identification, and estimation of these coefficients required a commitment to a particular parametric and distributional model and fell short of providing a general, causally defensible measure of mediation (Glynn, 2012; Hayes, 2009; Kraemer, Kiernan, Essex, & Kupfer, 2008; MacKinnon, 2008).

This has changed in the past 2 decades. Counterfactual thinking in statistics (Holland, 1988; Rubin, 1974) and epidemiology (Robins & Greenland, 1992), together with a formal semantics based on nonparametric structural equations (Balke & Pearl, 1995; Halpern, 1998; Pearl, 2001), has given causal mediation analysis a sound theoretical basis and extended its scope from linear to nonlinear models. The definitions of direct and indirect effects that emerge from this graphical-counterfactual symbiosis (summarized in the Natural Direct and Indirect Effects section, below) require no commitment to functional or distributional forms and are therefore applicable to models with arbitrary nonlinear interactions, arbitrary dependencies among the random variables, and both continuous and categorical variables.

This article concerns the conditions under which direct and indirect effects can be estimated from observational studies. In particular, I focus on the natural mediated effect, which is defined (roughly) as the expected change in the output when one lets the mediator change as if the input did (see the Natural Direct and
Indirect Effects section, below, for formal definition). This counterfactual entity, which has engendered the transition from linear to nonlinear models, cannot, in general, be estimated from controlled experiments, even when it is feasible to randomize both the treatment and the mediating variables.\textsuperscript{1} This limitation, noted by Robins and Greenland in 1992, resulted in 9 years of abandonment, during which natural effects were considered void of empirical content and were not investigated (S. Kaufman, Kaufman, & MacLenose, 2009).

Interest in natural effects rekindled when identification conditions were uncovered that circumvented this limitation, mediation formulas were derived, and the role of natural effects in policy making was made explicit (Pearl, 2001). While the identification conditions relied on untenable assumptions, those assumptions were conceptually meaningful and not substantially different from standard requirements of \textit{no confounding or no common causes} that are made routinely in causal analysis.\textsuperscript{2}

These developments, coupled with the capability of expressing and visualizing causal assumptions in graphical forms, have given rise to an explosion of mediation studies that have taken natural effects as the gold standard for analysis (e.g., Albert & Nelson, 2011; Coffman & Zhong, 2012; Hafeman & Schwartz, 2009; Huber, 2012; Imai, Keele, Tingley, & Yamamoto, 2011; Imai, Keele, & Yamamoto, 2010; Jo, Stuart, MacKinnon, & Vinokur, 2011; Joffe, Small, & Hsu, 2007; J. Kaufman, 2010; Mortensen, Diderichsen, Smith, & Andersen, 2009; Petersen, Sinisi, & van der Laan, 2006; Richardi, Bellocco, & Zuga, 2013; Robins, 2003; Sobel, 2008; Ten Have, Elliott, Joffe, Zanutto, & Datto, 2004; Valeri & VanderWeele, 2013; VanderWeele & Vansteelandt, 2009; Vansteelandt, Bekaert, & Lange, 2012). These studies have also adopted the mediation formulas of natural effects as targets for estimation and as benchmarks for sensitivity analysis (Imai, Keele, & Yamamoto, 2010; Sjölander, 2009).

However, although the identification conditions invoked in current mediation analysis are based on the same formal principles (see Appendix B),\textsuperscript{3} the articulation of these conditions in common scientific terms becomes highly varied and unreliable, making it hard for researchers to judge their plausibility in any given application. This stems from the difficulty of discerning conditional independencies among counterfactual variables, which must be undertaken by rank-and-file researchers whenever natural effects need be identified (Imai, Keele, & Yamamoto, 2010; Pearl, 2001; Petersen et al., 2006; Robins, 2003; VanderWeele & Vansteelandt, 2009). The verification of such independencies, often called \textit{strong ignorability, conditional ignorability, or sequential ignorability}, presents a formidable judgmental task to most researchers if unaided by structural models (Joffe, Yang, & Feldman, 2010).

Recently, efforts have been made to interpret these conditions in more conceptually meaningful way, so as to enable researchers to judge whether the necessary assumptions are scientifically plausible (Coffman & Zhong, 2012; Imai, Jo, & Stuart, 2011; Imai, Keele, & Tingley, 2010; Muthén, 2011; Richardi et al., 2013; Valeri & VanderWeele, 2013; VanderWeele, 2009). Invariably, these efforts strive to replace ignorability vocabulary with notions such as \textit{no unmeasured confounders}, \textit{no unmeasured confounding}, \textit{as if randomized}, \textit{effectively randomly assigned}, or \textit{essentially random}, which are clearly more meaningful to empirical researchers.

Unfortunately, these interpretations are marred by two sources of ambiguity. First, the notion of a confounder varies significantly from author to author. Some define a confounder (say, of X and Y) as a variable that affects both X and Y. Some define a confounder as a variable that is associated with both X and Y. Others allow for a confounder to affect X and be associated with Y. Worse yet, the expression \textit{no unmeasured confounders} is sometimes used to exclude the very existence of such confounders and sometimes to affirm our ability to neutralize them by controlling other variables, not necessarily confounders. Second, the interpretations have taken sequential ignorability as a starting point and consequently are overly stringent—sequential ignorability is a sufficient but not necessary condition for identifying natural effects. Weaker conditions can be articulated in a transparent and unambiguous language providing a greater identification power and a greater conceptual clarity.

A typical example of overly stringent conditions that can be found in the literature reads as follows:

The sequential ignorability assumption must be satisfied in order to identify the average mediation effects. This key assumption implies that the treatment assignment is essentially random after adjusting for observed pretreatment covariates and that the assignment of mediator values is also essentially random once both observed treatment and the same set of observed pretreatment covariates are adjusted for. (Imai, Jo, & Stuart, 2011, pp. 863–864)\textsuperscript{4}

I show that milder conditions are sufficient for identification. First, there is no need to require that covariates be pretreatment, as long as they are causally unaffected by the treatment. Second, the treatment assignment need not be random under any adjustment; identification can be achieved with treatment assignment remaining highly confounded under every set of observed covariates. Finally, one need not insist on using “the same set of observed pretreatment covariates”; two or three different sets can sometimes accomplish what the same set cannot.

On the other extreme, there is also a tendency among researchers to treat the necessary adjustments as totally independent of each other. A common misconception presumes that control of confounding between the treatment and the mediator can be accomplished independently of how one controls confounding between the mediator and the outcome. I show this not be the case;

\textsuperscript{1} This is because there is no way to rerun history and measure each subject’s response under conditions he or she has not actually experienced.

\textsuperscript{2} Discussion about the philosophical and practical implications of this limitation can be found in Pearl (2009b, pp. 35, 391) and Robins and Richardson (2011). The rest of the article assumes that the investigator is in possession of scientific knowledge to judge the plausibility of \textit{no confounding} type of assumptions that underlie all current research on mediation whether under the rubric of \textit{sequential ignorability} (e.g., Imai, Keele, & Yamamoto, 2010) or \textit{uncorrelated error terms}.

\textsuperscript{3} Imai, Keele, and Yamamoto (2010) and Imai, Keele, Tingley, and Yamamoto (2011) discussed similarities and differences among several versions of the identifying assumptions, and Shpitser and VanderWeele (2011) delineated the context under which a restricted version of the conditions established in Pearl (2001) coincide with those established in Imai, Keele, and Yamamoto.

\textsuperscript{4} A formal description of this and other identification strategies can be found in Imai, Keel, and Tingley (2010, Section 3.3) and Imai, Keele, Tingley, and Yamamoto (2011); the latter supplements the description with graphs to facilitate communication.
Mediation Analysis in the Parametric Tradition

In this section, I introduce mediation analysis from the perspective of nonparametric SEMs. This approach integrates the potential outcome framework of Splawa-Neyman (1923/1990) and Rubin (1974) with that of SEM, thus combining mathematical rigor with the merits of staying intimately informed by the data-generating process or its graphical representation.

Mediation Analysis in the Parametric Tradition

Figure 1 depicts the basic mediation structure that I later embed in wider contexts. It consists of three random variables: T, often called treatment; Y, the outcome; and M, the mediator, whose role in transmitting the effect of T on Y I wish to assess. As a running example, one could imagine an encouragement design (Holland, 1988) where T stands for a type of educational program that a student receives, M stands for a student’s score on the exam. In the linear case (see Figure 1a), the causal relationships in this example would be modeled in three linear equations:

\[
t = u_T, \quad m = \alpha t + u_M, \quad y = \beta t + \gamma m + u_Y,
\]

where lowercase symbols (t, m, y) represent the values that the variables (T, M, Y) may take and \( U_T, U_M, \) and \( U_Y \) stand for omitted factors that explain variations in T, M, and Y. The coefficients \( \alpha, \beta, \) and \( \gamma \) represent the structural parameters that need to be estimated from the data and that define the direct (\( \beta \)), indirect (\( \alpha\gamma \)) and total (\( \tau = \beta + \alpha\gamma \)) effects of T on Y.

As structural parameters, \( \alpha, \beta, \) and \( \gamma \) are causal quantities whose meaning is independent of the methods used in their estimation. \( \gamma \), for example, stands for the increase in a student’s score (Y) per unit increase in study time (M), keeping all other factors (T and \( U_Y \)) constant. This unit-based, ceteris paribus definition of structural parameters may lend itself to experimental verification when certain conditions hold. The assumption of linearity, for example, renders structural coefficients constant across individuals and permits one to estimate them by controlled experiments at the population level. One can imagine, for example, an investigator going to a district where T is not available, recruiting interested students (and their parents) and then randomly assigning T = 1 to some and T = 0 to others, and estimating \( \alpha \) through the difference in the mean of M between the two experimental groups, which we write as \( E[M \mid do(T = 1)] - E[M \mid do(T = 0)] \). At the same experiment, the investigator can also measure students’ scores, Y, and estimate the total effect

\[
\tau = E[Y \mid do(T = 1)] - E[Y \mid do(T = 0)] = \beta + \alpha \gamma.
\]

To estimate \( \gamma \) would require a more elaborate experiment in which both T and M are simultaneously randomized, thus disconfounding all three relationships in the model and permitting an unbiased estimate of \( \gamma \):

\[
\gamma = E[Y \mid do(T = 0), do(M + 1)] - E[Y \mid do(T = 0), do(M)].
\]

The latter can also be estimated in an encouragement design where M is controlled not directly but through a randomized incentive for homework. However, most traditional work on mediation focused on nonexperimental estimation, treating the structural equations in Equation 1 as regression equations, assuming that each U term is uncorrelated with the predictors in the same equation.

The regression analysis of mediation, most notably the one advanced by Baron and Kenny (1986), can be stated as follows: To test the contribution of a given mediator M to the effect of T on Y, first regress Y on T, and estimate the regression coefficient \( R_{YT} \), to be equated with the total effect \( \tau \). Second, include M in the regression, and estimate the partial regression coefficient \( R_{YT,M} \).

**Figure 1.** a: The basic (unconfounded) mediation model. b: A confounded version of a, showing correlation between \( U_M \) and \( U_Y \). Solid bullets represent observed variables; hollow circles represent unobserved (or latent) variables. M = mediator; T = treatment; U = omitted factors; Y = outcome; \( \alpha, \beta, \gamma = \) structural coefficients.
when $M$ is controlled for (or conditioned on or adjusted for). The difference between the two slopes, $R_{YX} - R_{YM}$, would then measure the reduction in the total effect due to controlling for $M$ and should quantify the effect mediated through $M$.

The rationale behind this estimation scheme follows from Figure 1a. If the total effect of $T$ on $Y$ through both pathways is $\tau = \beta + \alpha \gamma$, by adjusting for $M$, one severs the $M$-mediated path, and the effect is reduced to $\beta$. The difference between the two regression slopes gives the indirect, or mediated effect

$$\tau - \beta = \alpha \gamma. \quad (2)$$

Alternatively, one can venture to estimate $\alpha$ and $\gamma$ independently of $\tau$. This is done by first estimating the regression slope of $M$ on $T$ to get $\alpha$, then estimating the regression slope of $Y$ on $M$ controlling for $T$, which gives us $\gamma$; multiplying the two slopes together gives us the mediated effect $\alpha \gamma$.

The validity of these two estimation methods depends on course on the assumption that the error terms, $U_T$, $U_M$, and $U_Y$, are uncorrelated. Otherwise, some of the structural parameters might not be estimable by simple regression, and both the difference-in-coefficients and product-of-coefficients methods will produce biased results. In randomized trials, where $U_T$ can be identified with the randomized treatment assignment, we are assured that $U_T$ is uncorrelated with both $U_M$ and $U_Y$, so the regression estimates of $\tau$ and $\alpha$ will be unbiased. However, randomization does not remove correlations between $U_{X_T}$ and $U_X$. If such correlation exists (as depicted in Figure 1b), adjusting for $M$ will create spurious correlation between $T$ and $Y$, which will prevent the proper estimate of $\gamma$ or $\beta$. In other words, the regression coefficient $R_{XY}$ will no longer equal $\gamma$, and the difference $R_{XY} - R_{XM}R_{YM}$ will no longer equal $\beta$. This follows from the fact that controlling or adjusting for $M$ in the analysis (by including $M$ in the regression equation) does not physically disable the paths going through $M$; it merely matches samples with equal $M$ values and thus induces spurious correlations among other factors in the analysis (see Bullock, Green, & Ha, 2010; Cole & Hernán, 2002; Pearl, 1998; VanderWeele & Vansteelandt, 2009). Such correlations cannot be detected by statistical means, so theoretical knowledge must be invoked to identify the sources of these correlations and control for common causes (so called “confounders”) of $M$ and $Y$ whenever they are observable.8

This approach to mediation has two major drawbacks. One (mentioned above) is its reliance on the untested assumption of uncorrelated errors, and the second is its reliance on linearity and, in particular, on a property of linear systems called effect constancy (or no interaction): The effect of one variable on another is independent of the level at which we hold a third. This property does not extend to nonlinear systems; in such systems, the level at which we control $M$ would in general modify the effect of $T$ on $Y$. For example, if the output $Y$ requires both $T$ and $M$ to be present, then holding $M$ at zero would disable the effect of $T$ on $Y$, while holding $M$ at a high value would enable the latter.

As a consequence, additions and multiplications are not self-evident in nonlinear systems. It may not be appropriate, for example, to define the indirect effect in terms of the difference in the total effect, with and without control. Nor would it be appropriate to multiply the effect of $T$ on $M$ by that of $M$ on $Y$ (keeping $X$ at some level)—multiplicative compositions demand their justifications. Indeed, all attempts to define mediation by generalizing the difference and product strategies to nonlinear system have resulted in distorted and irreconcilable results (Glynn, 2012; MacKinnon, Fairchild, & Fritz, 2007; MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007; Pearl, 2012b).

The next section removes these nonlinear barriers by defining effect as a counterfactual notion, independent of any statistical or parametric manifestation, thus availing mediation analysis to a broad spectrum of new applications, primarily those involving categorical data and highly nonlinear processes. The first limitation, the requirement of error independence (or no unmeasured confounders), as it is often called is also relaxed, since the new definition opens new ways of overcoming correlations among the $U$ terms.

### Causes and Counterfactuals in Nonparametric Structural Models

In the most general case, the structural mediation model will have the form of Figure 2b:

$$t = f_T(u_T), \quad m = f_M(t, u_M), \quad y = f_Y(t, m, u_Y), \quad (3)$$

where $T, M, Y$ are discrete or continuous random variables, $f_T, f_M,$ and $f_Y$ are arbitrary functions, and $U_T$, $U_M$, and $U_Y$ represent, respectively, omitted factors that influence $T, M, Y$ but are not influenced by them. In our example, $U_m$ represents all factors that explain variations in study time ($M$) among students at the same treatment ($T$). The triplet $U = (U_T, U_M, U_Y)$ is a random vector that accounts for all variations between individual students. It is sometimes called unit, for it offers a complete characterization of a subject’s behavior as reflected in $T, M$, and $Y$. The distribution of $U$, denoted $P(U = u)$, uniquely determines the distribution $P(t, m, y)$ of the observed variables through the three functions in Equation 3.

In Figure 2a, the omitted factors are assumed to be arbitrarily distributed but mutually independent, written $U_T \perp \perp U_M \perp \perp U_Y$. In Figure 2b, the dashed arcs connecting $U_T$ and $U_M$ (as well as $U_M$ and $U_Y$) encode the understanding that the factors in question may be dependent. Figure 2c is a shorthand notation for Figure 2b. Here, the $U$ factors are not shown explicitly, and their dependencies are encoded in the form of dashed arcs going directly to the affected variables.

Referring to the student-encouragement example, it is not hard to imagine sources of possible dependencies among the omitted factors. For example, if $U_Y$ includes student’s intelligence and the amount of time studied varies systematically with intelligence, $U_M$ and $U_Y$ will be dependent, as shown in the model of Figure 2b. Likewise, if $U_T$ includes the propensity of students to enter the program ($T$) and this propensity depends on whether students have adequate conditions for...
home studies \((U_{M})\), then an arc between \(U_{T}\) and \(U_{M}\) is needed to encode their dependence (see Figure 2b). In general, as soon as one associates a diagram to a research context, interesting issues arise of possible associations among measured and unmeasured variables. Some can be decided by scientific considerations, and some may be debated by experts in the field. The purpose of the diagram is to provide an unambiguous description of the scientific context of a given application. While the application itself is usually shrouded in consensus on what is plausible and important versus that which is deemed negligible or implausible.

In this article, I emphasize the use of diagrams as faithful conveyors of the scientific context in any given application, with the understanding that the actual causal story behind the context may vary from problem to problem and that questions regarding the statistical and counterfactual implications of the diagrams can be answered mechanically by simple path-tracing routines.\(^9\) Notably, a model like that shown in Figure 2c allows for the existence of millions of unobserved subprocesses that make up the functions \(f_{T}, f_{M},\) and \(f_{Y}\); these do not alter questions concerning the mediating role of \(M\).

Since every SEM can be translated into an equivalent counterfactual (or potential outcome) model (Pearl, 2009b, Definition 7.1.5), we can give the mediation model of Equation 3 a counterfactual interpretation as follows. Define the counterfactual variables \(M_{t}, Y_{t},\) and \(Y_{m}\) by

\[
M_{t} = f_{M}(t, U), \quad Y_{t} = f_{Y}(t, M_{t}, U), \quad Y_{m} = f_{Y}(t, m, U),
\]

where \(U = (U_{T}, U_{M}, U_{Y})\) is the random variable representing all omitted factors. In other words, the counterfactual variable \(M_{t}\) stands for the value that \(M\) would take when we set the subscripted variable \(T\) to a constant \(t\) and allow the other variables in the equation (i.e., \(U\)) to vary. Similarly, \(Y_{m}\) stands for the value that \(Y\) would take when we set the subscripted variables \(T\) and \(M\) to constants, \(t\) and \(m\), and allow \(U\) to vary. Accordingly, the independence assumption \(U_{T} \perp (U_{M}, U_{Y})\) depicted in Figures 1b and 2a can be given a counterfactual form (called treatment ignorability):

\[
T \perp (M_{t}, Y_{t,m}) \quad \text{for all } t \text{ and } t',
\]

while \((U_{T}, U_{M}) \perp U_{Y}\) (depicted in Figure 2a) conveys the independence:

\[
(T, M_{t}) \perp Y_{t,m} \quad \text{for all } t \text{ and } t'.
\]

This translation from independence of omitted factors into independence of counterfactuals reflects the fact that the statistical variations of \(Y_{m}\) are caused solely by variations in \(U_{Y}\), since \(t\) and \(m\) are constants, and similarly, variations of \(M_{t}\) are caused solely by those of \(U_{M}\).

Since the functions \(f_{T}, f_{M},\) and \(f_{Y}\) are unknown to investigators, mediation analysis commences by first defining total, direct, and indirect effects in terms of those functions and then asking whether they can be expressed in terms of the available data, which we assume are given in the form of random samples \((t, m, y)\) drawn from the joint probability distribution \(P(t, m, y)\). Whenever such a translation is feasible, we say that the respective effect is identifiable.

**Natural Direct and Indirect Effects**

Using the structural model of Equation 3, four types of effects can be defined for the transition from \(T = 0\) to \(T = 1\).\(^{11}\)

**Total effect (TE).**

\[
TE = E[f_{Y}(1, f_{M}(1, U_{M}), U_{Y})] - E[f_{Y}(0, f_{M}(0, U_{M}), U_{Y})] = E[Y_{T=1} - Y_{T=0}].
\]

\[
TE = E[Y|do(T = 1)] - E[Y|do(T = 0)].
\]

**Controlled direct effect (CDE).**

\[
CDE(m) = E[f_{Y}(1, M = m, U_{Y})] - E[f_{Y}(0, M = m, U_{Y})] = E[Y_{1,m} - Y_{0,m}].
\]

\[
CDE(m) = E[Y|do(T = 1, M = m)] - E[Y|do(T = 0, M = m)].
\]

\(^9\) Readers who wish to read the statistical dependencies that a given context entails are advised to do so through the tool of \(d\)-separation (gently introduced in Appendix A), but this is not absolutely necessary, since \(d\)-separation and other graph-based techniques are mechanized on several available software programs (e.g., Kyono, 2010; Textor, Hardt, & Knüppel, 2011).

\(^{11}\) Generalizations to arbitrary reference point, say, from \(T = t\) to \(T = t'\), are straightforward. These definitions apply at the population levels; the unit-level effects are given by the expressions under the expectation. All expectations are taken over the factors \(U_{M}\) and \(U_{Y}\).
\( CDE \) measures the expected increase in the outcome \( Y \) as the treatment changes from \( T = 0 \) to \( T = 1 \), while the mediator is set to a prespecified level \( M = m \) uniformly over the entire population.

**Natural direct effect (NDE)**

\[
NDE = E[f_y(1, f_M(0, u_M), u_f)] - f_y[0, f_M(0, u_M), u_f]] = E[Y_{1,m} - Y_{0,m}].
\] (9)

NDE measures the expected increase in \( Y \) as the treatment changes from \( T = 0 \) to \( T = 1 \), while setting the mediator variable to whatever value it would have attained (for each individual) prior to the change, that is, under \( T = 0 \).

**Natural indirect effect (NIE)**

\[
NIE = E[f_y(0, f_M(1, u_M), u_f)] - f_y[0, f_M(0, u_M), u_f]] = E[Y_{0,M_1} - Y_{0,M_0}].
\] (10)

NIE measures the expected increase in \( Y \) when the treatment is held constant, at \( T = 0 \), and \( M \) changes to whatever value it would have attained (for each individual) under \( T = 1 \).

Semantically, \( NDE \) measures the portion of the total effect that would be transmitted to \( Y \) absent \( M \)’s ability to respond to \( T \), while \( NIE \) measures the portion transmitted absent \( Y \)’s ability to respond to changes in \( T \), except those transmitted through \( M \). The difference \( TE - NDE \) quantifies the extent to which the response of \( Y \) is owed to mediation, while \( NIE \) quantifies the extent to which it is explained by mediation. These two components of mediation, the necessary and the sufficient, coincide into one in models void of interactions (e.g., linear) but differ substantially under moderation (see the Numerical Example section, below).

We remark that a controlled version of \( NIE \) does not exist because there is no way of disabling the direct effect of \( T \) on \( Y \) by setting a variable to a constant. Note also that the natural effects, \( NDE \) and \( NIE \), are not accompanied by \( do \)-expressions because these effects are defined counterfactually and cannot be estimated from controlled experiments. The choice of the appropriate effect type in policy making is discussed in Pearl (2001, 2011), Robins and Richardson (2011), and VanderWeele (2009) and are illustrated in the Illustrations section, below. Whereas the controlled direct effect is of interest when policy options exert control over values of \( variables \) (e.g., raising the level of a substance in patients’ blood to a prespecified concentration), the natural direct effect is of interest when policy options enhance or weaken mechanisms or processes (e.g., freezing a substance at its current level of concentration [for each patient], but preventing it from responding to a given stimulus).

This is an appropriate point to relate the definitions of natural effects to the standard definitions of direct and indirect effects used in parametric structural equation. When we apply the definitions above to the linear system of Equation 1, we readily obtain the expected results:

\[
TE = \beta + \alpha \gamma, \quad NDE = CDE(m) = \beta, \quad NIE = \alpha \gamma.
\] (11)

A key conceptual difference between the causal and the traditional approaches is that, in the former, every effect is defined a priori, in a way that makes it applicable to any model, including confounded, unidentified, or nonlinear models. The statistical approach, on the other hand, requires that the model satisfies certain restrictions before the definition (of effects) obtains its legitimacy. This is somewhat paradoxical, for one must know what one seeks to estimate before imposing the appropriate restrictions on the model.

The equalities in Equation 10, for example, are derived from the basic definitions of Equations 6–9 and the linearity of Equation 1; they are sustained therefore in all linear systems, even when one does not make the assumption of no omitted variables (or ignorability). Likewise, the constancy of the controlled direct effect in linear system, \( CDE(m) = \alpha \), is not an assumption but a consequence of how \( CDE(m) \) is defined (see Equation 8).

In the classical approach, on the other hand, the assumption of no omitted variables must precede all definitions (Judd & Kenny, 1981, 2010) because the classical vocabulary was restricted to the statistical notion of controlling for \( M \) instead of the intended causal notion of setting \( M \) to a constant, and the two coincide only under the no omitted variables assumption.\(^{12}\) (See Bollen & Pearl, 2013, for further discussion of this important observation, which is often overlooked in the potential-outcome literature; e.g., Rubin, 2010; Sobel, 2008.)

Finally, note that, in general, the total effect can be decomposed as

\[
TE = NDE - NIE,
\] (12)

where \( NIE \) stands for the natural indirect effect under the reverse transition, from \( T = 1 \) to \( T = 0 \). This implies that \( NIE \) is identifiable whenever \( NDE \) and \( TE \) are identifiable. In linear systems, where reversal of transitions amounts to negating the signs of their effects, one has the standard additive formula, \( TE = NDE - NIE \). Moreover, since each term in Equation 12 is based on an independent operational definition, this equality constitutes a formal justification for the additive formula taken for granted in linear systems.\(^{13}\)

**The Counterfactual Derivation of Natural Effects**

To make this article self-contained, Appendix B provides a formal proof of the conditions for direct effect identification, as it appeared in Pearl (2001). It starts with the counterfactual definition of the natural direct effect and then goes through three steps. First, it seeks a set of covariates \( W \) that reduces nested counterfactuals to simple counterfactuals. Second, it reduces all counterfactuals to \( do \)-expressions, that is, expressions that are estimable from controlled randomized experiments. Finally, it poses conditions for identifying the \( do \)-expressions from observational studies. These three steps are echoed in the informal conditions articulated in the next section. (See also Shpitser & VanderWeele, 2011, and especially Shpitser, 2013, for refinements and elaborations.)

---

\(^{12}\) It is interesting to note that Equation 10 remains valid under temporal reversal of the \( T \rightarrow M \) relationship, that is, \( \alpha = 0 \) and \( \beta = \delta M + U_y \). In such a model, Equations 7–10 give the correct result; \( TE = NDE = CDE = \beta \), \( NIE = 0 \). The statistical definition, on the other hand, with its vocabulary confined to regression slopes, would not recognize \( NIE \) as zero because the regression slope of \( M \) on \( T \) is nonzero.

\(^{13}\) Some authors (e.g., VanderWeele, 2009; Vansteelandt, 2012, Chapter 4.4), take \( NIE = TE - NDE \) as the definition of the natural indirect effect, which ensures additivity a priori but presents a problem of interpretation; the resulting indirect effect, aside from being redundant, does not represent the same direction of change, from \( T = 0 \) to \( T = 1 \), as do the total and direct effects. This makes it hard to compare the effect attributed to mediating paths with that attributed to unmediated paths under the same conditions of change.
Interpretable Conditions for Identification

Preliminary Notation and Nomenclature

In this section, I provide precise identification conditions based solely on the notion of unconfoundedness. I say that the relationship between T and Y is unconfounded if the factors that influence T are independent of all factors that influence Y when T is held fixed. Given a set W of covariates, I say that W renders a relationship unconfounded if the relationship is unconfounded in every stratum W = w of W. Finally, I use the expression W deconfounds a relationship as a shorthand substitute for W renders a relationship unconfounded. This definition also provides a model-based interpretation of conditional strong ignorability, written T \perp\!\!\!\!\!\!\!\!\!\perp (Y, Y_0) | W, and can be given a simple graphical representation called backdoor (see Appendix A), as is illustrated in the next section. Deconfounding occurs, for example, if W consists of all common causes of T and Y but may hold for other types of covariates as well (known as sufficient or admissible; see Appendix A), which neutralize the effect of common causes. Figuratively, such deconfounders can be recognized by intercepting, or blocking, all spurious (noncausal) paths between T and Y, namely, all paths that end with an arrow toward T (also called backdoor paths).14

In Figure 1a, for example, the relationship between M and Y is confounded by T, the common cause of M and Y. T is also a deconfounder of this relationship because T blocks the (one and only) backdoor path between M and Y. In Figure 1b, on the other hand, the relationship between M and Y is confounded by T as well as by latent common causes represented by the dashed arc between M and T. Again, however, any measured variable. However, if U_M were to be observed, then the set W = [T, U_M] (similarly W = [T, U_J]) would deconfound the M \rightarrow Y relationship by blocking all backdoor paths from M to Y. Note that U_M in this case is a deconfounder though it is not a common cause of M and Y.

I focus my discussion on the natural direct effect, NDE, though all conditions are applicable to the indirect effect as well, by virtue of the pseudoadditive decomposition of the total effect (see Equation 12). I assume that readers are familiar with the notion of identifiability as applied to causal or counterfactual relations (see, e.g., Appendix A). In particular, I say that the W-specific causal effect of T on Y is identifiable if the effect is consistently estimable from noneperimental data for every stratum level w. In other words, the causal effect P(y \mid do(t, w)) can be expressed in terms of conditional probabilities of observed variables.15 It is important to note that the problem of deciding whether such reduction exists has been fully solved using the do-calculus (Shpitser & Pearl, 2008; Tian & Shpitser, 2010). Consequently, effective algorithms are available that, given any causal diagram, can reduce any do-expression—in particular, TE, CDE(m), and P(y \mid do(t_1, t_2, \ldots, t_k), (w_1, w_2, \ldots, w_j))—to regression expressions, whenever such reduction exists. I therefore regard the identifiability of do-expressions as a solved problem and focus my attention on the question of whether NDE and NIE can be thus expressed and how.

Sufficient Conditions for Identifying Natural Effects

The following are two sets of assumptions or conditions, marked A and B, that are sufficient for identifying both direct and indirect natural effects. Each condition is communicated by a verbal description followed by its formal expression. Each set of conditions is followed by its graphical version, marked A_G and B_G, with all graphs representing nonparametric SEMs,16 as in Figure 2. Assumption Set B is the stronger of the two and represents assumptions commonly invoked in the mediation literature (Coffman & Zhong, 2012; Imai, Jo, & Stuart, 2011; Imai, Keele, & Yamamoto, 2010; Shpitser & VanderWeele, 2011; VanderWeele & Vansteelandt, 2009; Vansteelandt et al., 2012; Vansteelandt & Lange, 2012). Assumption Set A is weaker and echoes more faithfully the derivation in Appendix B. For completeness, I also present a third assumption set, C, representing a compromise between A and B, which is based solely on the presence of deconfounding covariates, thus echoing more closely the way assumptions are articulated in the literature (e.g., Valeri & VanderWeele, 2013). Following a listing of the three assumption sets, Theorem 1 then presents the general formula for the natural direct effect (NDE) that results from Assumption Set A. The corresponding formula that results from Assumption Set B is given in Corollary 2. The corresponding formulas for the NIE follow from Equation 12 and are explicated in Equation 14b.

Assumption Set A. There exists a set W of measured covariates such that

A-1. No member of W is affected by treatment;

\[ M_t \perp\!\!\!\!\!\!\!\!\!\perp Y_{t,w} \mid W \] (alternatively, \[ U_M \perp\!\!\!\!\!\!\!\!\!\perp U_I \mid W \]);

A-3. The W-specific effect of the treatment on the mediator is identifiable by some means:

\[ P(m \mid do(t, w)) \text{ is identifiable}; \]

A-4. The W-specific joint effect of \( (treatment + mediator) \) on the outcome is identifiable by some means:

\[ P(y \mid do(t, m), w) \text{ is identifiable}. \]

14 By path, I mean any sequence of adjacent edges, regardless of directionality. By blocking, I mean disconnecting the path in the d-separation sense (see Appendix A).  
15 The expression P(y \mid do(t, w)) stands for the conditional probability P(y \mid T = t, W = w) obtained in a controlled experiment in which T is randomized and in which only units for which W = w are recorded. TE and CDE(m) are do-expressions and can, therefore, be estimated from experimental data; not so the natural effects, NDE and NIE can be estimated from experimental data only when additional no confounding conditions hold (see Footnote 11) to be explicated below. The do-calculus (Pearl, 1995, 2009b, pp. 85–88) is a method of systematically reducing do-expressions to ordinary conditional probabilities but is not needed in this article.  
16 The distinction between graphs representing SEMs versus interventional models is discussed at length in Pearl (2009b, pp. 22–38) and is further elaborated in Robins and Richardson (2011). The latter models are also known as causal Bayesian networks; they represent experimental findings (i.e., do-expressions) but do not sanction counterfactual inferences.
Graphical version of Assumption Set A. There exists a set $W$ of measured covariates such that

- **A$_C$-1.** No member of $W$ is a descendant of $T$;
- **A$_C$-2.** $W$ blocks all backdoor paths from $M$ to $Y$ not traversing $T$;
- **A$_C$-3.** The $W$-specific effect of $T$ on $M$ is identifiable (possibly using auxiliary variables); and
- **A$_C$-4.** The $W$-specific joint effect of $\{T, M\}$ on $Y$ is identifiable (possibly using auxiliary variables).

Illustration of $A_{C}$. Figure 3a provides an example where all $A_{C}$ conditions are satisfied by $W = W_{1}$. First, $W_{1}$ satisfies $A_{C}$-1 and $A_{C}$-2 by virtue of being a non-descendant of $T$ and blocking the path $M \leftarrow W_{1} \rightarrow Y$, the only backdoor path from $M$ to $Y$ that does not traverse $T \rightarrow M$ or $T \rightarrow Y$ or that is not already blocked (by $\{W\}$). Next, $A_{C}$-3 is satisfied because the set $\{W_{1}, W_{2}\}$ deconfounds the $T \rightarrow M$ relationship. This renders the $W_{1}$-specific causal effect of $T$ on $M$ identifiable by adjusting for $W_{2}$ and yields $P(y | do(t, w_{1})) = \sum_{w_{2}} P(y | t, w_{2}, w_{1}) P(w_{2})$. The same applies to $A_{C}$-4, using adjustment for $W_{3}$ to identify the $W_{3}$-specific effect of $\{T, M\}$ on $Y$, yielding $P(y | do(t, m), w_{1}) = \sum_{w_{2}} P(y | t, m, w_{2}, w_{1}) P(w_{2})$.

Assumption Set B (sequential ignorability, Imai, Keele, & Yamamoto, 2010). There exists a set $W$ of measured covariates such that

- **B-1.** $W$ and $T$ deconfound the mediator-outcome relationship, keeping $T$ fixed:
  
  $$[Y_{f,m} \perp M_{f} | T, W];$$  
  and

- **B-2.** $W$ deconfounds the treatment-[mediator, outcome] relationship:
  
  $$[T \perp (Y_{f,m}, M_{f}) | W].$$

Graphical version of Assumption Set B. There exists a set $W$ of measured covariates such that

- **B$_C$-1.** $W$ and $T$ block all $T$-avoiding backdoor paths from $M$ to $Y$; and
- **B$_C$-2.** $W$ blocks all backdoor paths from $T$ to $M$ or to $Y$, and no member of $W$ is a descendant of $T$.

Illustration of $B_{C}$. Figure 3a provides an example where all $B_{C}$ conditions are satisfied using $W = \{W_{1}, W_{2}, W_{3}\}$. First, we examine all $T$-avoiding backdoor paths from $M$ to $Y$ (in particular, $M \leftarrow W_{1} \rightarrow Y$) and note that $\{W_{1}, W_{2}, W_{3}\}$ blocks those paths, thus satisfying $B_{C}$-1. Next, complying with $B_{C}$-2, the set $\{W_{1}, W_{2}, W_{3}\}$ blocks the paths $T \leftarrow W_{1} \rightarrow M$ and $T \leftarrow W_{2} \rightarrow Y$, the only ones with arrows into $T$. Finally, none of $W_{1}, W_{2}, W_{3}$ is a descendant of $T$, thus satisfying $B_{C}$-2.

Note that conditions $A$-3 and $B$-2 are automatically satisfied if $T$ is randomized and $A$-4 is satisfied when both $T$ and $M$ are randomized, but the same is not true of $A$-2 and $B$-1; these may not hold even when we randomize both $T$ and $M$ (see Footnote 1).

If we limit the identification conditions to only those that invoke adjustment for covariates (giving up the options of using more elaborate identification methods, as in $A$-3 and $A$-4) Assumption Set A can be articulated more concisely thus:

Assumption Set C (piecemeal deconfounding). There exists three sets of measured covariates $W = \{W_{1}, W_{2}, W_{3}\}$ such that

- **C-1.** No member of $W_{1}$ is affected by the treatment;
- **C-2.** $W_{1}$ deconfounds the $M \rightarrow Y$ relationship (holding $T$ constant);
- **C-3.** $\{W_{2}, W_{1}\}$ deconfounds the $T \rightarrow M$ relationship; and
- **C-4.** $\{W_{3}, W_{1}\}$ deconfounds the $\{T, M\} \rightarrow Y$ relationship.

Note that $C$-4 is sufficient for identifying the controlled direct effect (see Equation 8), $C$-3 and $C$-4 are sufficient for identifying the total effect (see Equation 7), and all four conditions are needed for the natural effects.

Theorem 1 (Pearl, 2001): When Conditions A-1 through A-2 hold, the natural direct effect is identified and is given by

$$NDE = \sum_{m} \sum_{w} \{E(Y | do(T = 1, M = m), W = w) - E(Y | do(T = 0, M = m), W = w)\} P(M = m | do(T = 0), W = w)P(W = w).$$

(13)

Corollary 1: If Conditions A-1 and A-2 are satisfied by a set $W$ that also deconfounds the relationships in A-3 and A-4, then the $do$-expressions in Equation 13 are reducible to conditional expectations, and the natural direct and indirect effects become

---

17 This provision reflects the constancy of $T$ in Assumption A-2 as depicted in Figure 2b. Both $U_{m}$ and $U_{y}$ are defined relative to the condition where $T$ is held constant, a condition that precludes $T$ from passing information (or creating dependencies) between $U_{m}$ and $Y_{y}$.

18 Summations should be replaced by integration when applied to continuous variables, as in Imai, Keele, and Yamamoto (2010). Note that Equation 13 is still valid if only A-1 and A-2 are satisfied by $W$. A-3 and A-4 are needed solely for identifying the $do$-expressions in the equation.

19 Equations 14a-14b are identical to the ones derived by Imai, Keele, and Yamamoto (2010) using sequential ignorability (i.e., Assumptions B-1 and B-2) and subsequently rederived by a number of other authors (Lindquist, 2012; Wang & Sobel, 2013).
$NDE = \sum_m \sum_w [E(Y \mid T = 1, M = m, W = w) - E(Y \mid T = 0, M = m, W = w)]$ 

(14a)

$P(M = m \mid T = 0, W = w)P(W = w)$.

(14b)

$NIE = \sum_m \sum_w [P(M = m \mid T = 1, W = w) - P(M = m \mid T = 0, W = w)]E(Y \mid T = 0, M = m, W = w)$.

Equations 14a and 14b are the averages (over $w$) of the mediation formula (i.e., Equations 17 and 27 in Pearl, 2001; see Footnote 20 below) and were called the adjustment formula in Shpitser and VanderWeele (2011).

**Corollary 2:** If conditions $B$-1 and $B$-2 are satisfied by a set $W$, then the natural direct and indirect effects are identified and are given by Equations 14a and 14b.

Corollary 2 follows from Theorem 1 by noting that, in structural models, any set $W$ that satisfies $B$-1 and $B$-2 also deconfounds the relationships in $A$-3 and $A$-4 (Shpitser & VanderWeele, 2011).

**Corollary 3:** If Conditions $A$-1 and $A$-2 are satisfied with $W = \{2\}$ and two other sets of covariates exist, $W_2$ and $W_3$, such that $W_2$ deconfounds the $T \rightarrow M$ relationship and $W_3$ deconfounds the $\{TM\} \rightarrow Y$ relationship, then, regardless of possible dependencies between $W_2$ and $W_3$, the natural direct effect is identified and is given by

$$NDE = \sum_m \sum_{w_2} [E(Y \mid T = 1, M = m, W_3 = w_3) - E(Y \mid T = 0, M = m, W_3 = w_3)]P(W_3 = w_3)$$

$$\sum_{w_2} P(M = m \mid T = 0, W = w_2)P(W = w_2).$$

(15)

**Remarks.** Assumption Set $A$ differs from Assumption Set $B$ on two main provisions. First, $A$-3 and $A$-4 permit the identification of these causal effects by all methods, while $B$-2 and $B$-3 insist that identification be accomplished by adjustment. Second, whereas $A$-3 and $A$-4 allow for the invocation of any set of covariates in order to identify the $W$-specific effect in question, $B$ requires that the same set $W$ of covariates deconfound both the mediator-outcome and treatment-[mediator, outcome] relationships.

It should be noted that, whereas this article concerns identification in observational studies, Conditions $A$-3 and $A$-4 open the door to experimental studies, when such are feasible. For example, one may venture to estimate the causal effect of $T$ on $M$ by randomizing $T$ or by using instrumental variables or auxiliary intermediate variables. Only the latter are considered here. The restrictions on all such designs are the same, namely, that they lead to the identification of $W$-specific effects, where $W$ is a set of attributes satisfying $A$-1 and $A$-2. Assumption $A$-2, on the other hand, cannot be satisfied by any experimental design since it involves cross-world independence, from $t$-worlds to $t'$-worlds. Identifiability requires that such independencies hold naturally in the population under study, not in a population crafted by design (see Footnote 1).

**Appendix C** explains why I must insist that $W$ be unaffected by the treatment. This requirement is implicit in $B$-2 but not in $A$-2; it must therefore be stated explicitly in $A$-1 (and $B_{C2}$ for, otherwise, $A$-3 and $A$-4 will not be sufficient for identifying $NDE$, as is shown below.

### Illustrations

To illustrate and compare the conditions articulated in the previous section, I start with simple models that satisfy the strong conditions of $B$ (and $B_{C2}$), and then examine how the process of identification can benefit from the relaxed conditions given in $A$ and $A_{C3}$.

### How the Natural Effects Are Identified

Figure 4a illustrates the classical mediation model, with no confounding; all omitted factors (not shown in the diagram) affecting $T$, $M$, and $Y$ are assumed to be independent, so both the mediator process, $T \rightarrow M$, and the outcome process, $(T, M) \rightarrow Y$, are unconfounded. In this model, the null set $W = \{2\}$ satisfies the conditions in $B$ (as well as in $A$), and Equations 13 and 14a are reduced to

$$NDE = \sum_m [E(Y \mid T = 1, M = m) - E(Y \mid T = 0, M = m)]P(M = m \mid T = 0).$$

(16)

Likewise, the natural indirect effect (see Equation 14a) becomes

$$NIE = \sum_m E(Y \mid T = 0, M = m)$$

$$[P(M = m \mid T = 1) - P(M = m \mid T = 0)].$$

(17)

The intuition behind Equation 16 is simple; the natural direct effect is the weighted average of the controlled direct effect $CDE(m)$, shown in the square brackets, using the no-treatment distribution $P(M = m \mid T = 0)$ as a weighting function. Equation 16 can be estimated by a two-step regression, as is shown below. The intuition behind Equation 17 is somewhat different and unveils a nonparametric version of the product-of-coefficients estimator (see the Mediation Analysis in the Parametric Tradition section, above). The term $E(Y \mid T = 0, M = m)$ plays the role of $\gamma$ in Figure 1a, for it describes how $Y$ responds to $M$ for fixed treatment condition $(T = 0)$. The term in the square brackets plays the role of $\alpha$, for it captures the impact of the transition from $T = 0$ to $T = 1$ on the probability of $M$. One sees that what was a simple product operation in linear systems is replaced by a composition operator that involves summation over all values of $M$ and thus allows for heterogeneous populations where both $M$ and its effect on $Y$ may vary from individual to individual.

Figure 4b illustrates a confounded mediation model in which a variable, $W$ (or a set of variables), confounds all three relationships

---

20 Equations 16 and 17 were called the mediation formula in Pearl (2009b, p. 132; see also Pearl, 2009a, 2012a). Since the $NDE$ and $NIE$ are connected to each other via Equation 12, all our discussions concerning the identification of $NDE$ should apply to $NIE$ as well (Pearl, 2009b, p. 132; see also Pearl, 2009a, 2012a).
in the model. Because $W$ is not affected by $T$ and is observed, adjusting for $W$ renders all relationships unconfounded, and the conditions of $B$ (as well as $A$) are satisfied. Accordingly, the natural direct effect estimand is given by Equation 14b, which invokes the mediation formula (see Equation 16) in each stratum of $w$ of $W$. averaged over $w$.

### Numerical Example

To anchor these mediation formulas in a concrete example, I return to the encouragement-design example of the introduction and assume that $T = 1$ stands for participation in an enhanced training program, $Y = 1$ for passing the exam, and $M = 1$ for a student spending more than 3 hours per week on homework. Assume further that the data described in Table 1 were obtained in a randomized trial with no mediator-to-outcome confounding (see Figure 4a). The data show that training tends to increase both the time spent on homework and the rate of success on the exam. Figure 4b). The data show that training tends to increase both the time spent on homework and the rate of success on the exam.

Our research question asks for the extent to which students’ homework contributes to their increased success rates. The policy implications of such questions lie in evaluating policy options that either curtail or enhance homework efforts, for example, by counting homework effort in the final grade or by providing students with adequate work environments at home. An extreme explanation of the data, with significant impact on educational policy, might argue that the program does not contribute substantively to students’ success, save for encouraging students to spend more time on homework, an encouragement that could be obtained through less expensive means. Opposing this theory, there may be teachers who argue that the program’s success is substantive, achieved mainly due to the unique features of the curriculum covered, while the increase in homework efforts, although catalytical, cannot alone account for the success observed.

Substituting these data into Equations 16–17 gives

\[
NDE = (0.40 - 0.20)(1 - 0.40) + (0.80 - 0.30)0.40 = 0.32, \\
NIE = (0.75 - 0.40)(0.30 - 0.20) = 0.035, \\
TE = 0.80 	imes 0.75 + 0.40 	imes 0.25 - (0.30 	imes 0.40 + 0.20 	imes 0.10) = 0.46, \\
NIE/TE = 0.07, \quad NDE/TE = 0.696, \quad 1 - NDE/TE = 0.304.
\]

In conclusion, the program as a whole has increased the success rate by 46% and that a significant portion, 30.4%, of this increase is due to the capacity of the program to stimulate improved homework effort. At the same time, only 7% of the increase can be explained by stimulated homework alone without the benefit of the program itself.

Let me now illustrate the use of Equation 14a in cases marred by confounding. Assume that $W$ stands for gender, which, as shown in Figure 4b, confounds all three relations in the models. Equation 14a instructs us to conduct the analysis separately on males ($W = 1$) and females ($W = 0$) and average the results according to the gender mix in the population. For example, if the data in Table 1 represent the male population and a similar yet different table represents females, we take our estimate $NDE(W = 1) = 0.32$ and the corresponding $NDE(W = 0)$ from the female table and form the overall $NDE$ by taking the weighted average of the two.

The purpose of this example is to demonstrate how the linear barriers that restricted classical mediation analysis can be broken by nonparametric formulas, Equations 16 and 17, that have emerged from the structural-counterfactual analysis. It shows how these mediation formulas are applicable to highly interacting variables, both continuous and categorical, without making any assumptions about the error distribution or about the functions that tie the variables together. Imai, Keele, and Yamamoto (2010) further analyzed the asymptotic variance of the estimands in Equations 16 and 17 and developed powerful software for sensitivity analysis.

In the next section, I deal with more intricate patterns of confounders, both measured and unmeasured, and show how Conditions $A_{C-1}$ to $A_{C-4}$ can guide us toward identification in the presence of those confounders.

### The Benefits of Independent Adjustments

A benefit of the weaker conditions expressed in $A$ is that $A-3$ and $A-4$ allow for covariates outside $W$ to assist in the identification. This results in a greater flexibility in allocating covariates for the various adjustments invoked in Equation 14a. It also simplifies the process of justifying the assumptions that support these adjustments and leads, in turn, to a simpler overall estimand. Specifically, in choosing covariates to deconfound the $(T, M) \rightarrow Y$ relationship, one is free to ignore those chosen to deconfound the $T \rightarrow M$ relationship.

The model in Figure 3a demonstrates this flexibility. Although the set $W = \{W_1, W_2, W_3\}$ satisfies all the conditions in $A$ and $B$, Assumption Set $A$ permits us to handle each of the three covariates individually, so as to simplify the resulting estimand. Since $W_1$ alone renders the mediator-to-outcome relationship unconfounded.

### Table 1

<table>
<thead>
<tr>
<th>Treatment $T$</th>
<th>Homework $M$</th>
<th>$E(Y \mid T = t, M = m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.40</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$E(M \mid T = t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.40</td>
</tr>
<tr>
<td>0.75</td>
</tr>
</tbody>
</table>
(for fixed $T$), we are at liberty to choose $W_i$ to satisfy Conditions A-1 and A-2. In the next step, we seek a set of covariates that, together with $W_i$, would deconfound the $T \rightarrow M$ relationship, and since $W_i$ alone meets this requirement, we can remove $W_i$ from the factor $P(M = m \mid T = 0, W = w) = P(M = m \mid T = 0, W = w_i, W_2 = w_2, W_3 = w_3)$ of Equation 14a. Next, we seek a set of covariates that, together with $W_i$, would deconfound the $(T, M) \rightarrow Y$ relationship, and realizing that $W_i$ meets this requirement, we can remove $W_i$ from the factors $E(Y \mid T = 1, M = m, W = w)$ and $E(Y \mid T = 0, M = m, W = w)$ of Equation 14a. The resulting estimand for $NDE$ becomes

$$NDE = \sum_{m} \sum_{w_2, w_3, w_1} P(W_2 = w_2, W_3 = w_3, W_1 = w_1) \times P(M = m \mid T = 0, W_2 = w_2, W_1 = w_1) \times E(Y \mid T = 1, M = m, W = w, W_3 = w_3) - E(Y \mid T = 0, M = m, W = w, W_1 = w),$$

with only one of $W_3$ and $W_2$ appearing in each of the last two factors.

Note that covariates need not be pretreatment to ensure identification; $B$ and $A$ require merely that $W$ be causally unaffected by the treatment. Indeed, $W_i$ in Figure 3 may well be a posttreatment variable, the control of which is essential for identifying $NDE$.

Figure 3b associates a research context to the model of Figure 4a using our running example of student-encouragement design. Here, we assume that $W_1 = \text{reading skill}$ is the sole confounder of the homework $\rightarrow$ score relation. Likewise, we assume that socioeconomic background confounds program $(T)$ and homework $(M)$ ostensibly because students from high socioeconomic backgrounds are more likely to have facilities that are conducive to doing homework and they (or their parents) are more likely to seek out the educational programs offered $(T)$. Finally, we associate $W_3$ with students’ natural intelligence, arguing that this is a significant factor in enticing students to enroll in the program $(T)$ and simultaneously enables students to learn faster and score higher on exams.

As mentioned in the Structural Approach to Mediation section, above, as soon as one associates a diagram to a research context, issues arise of possible unforeseen associations among variables that may threaten identification and complicate estimation. In our example, mutual associations may naturally be suspected among language skills $(W_1)$, socioeconomic background $(W_2)$, and intelligence $(W_3)$, with no clear origin or explanation. Such associations are depicted by the dashed arcs in Figure 3b, and the question arises, Do these present a problem to identification? Such questions can be readily answered by Assumption Set A, using $A_{C\rightarrow-1}$ to $A_{C\rightarrow-4}$, though it is a bit hard to imagine how they can be handled by Assumption Set B.

Guided by $A_{C}$, note that all arguments previously used in deciding the identification of $NDE$ in Figure 3a (see the Sufficient Conditions for Identifying Natural Effects section, Illustration of $A_{C}$ subsection, above) are still valid for Figure 3b. Specifically,

(i) $W_i$ satisfies $A_{C\rightarrow-2}$ by virtue of blocking the two backdoor paths going from $M$ to $Y$, $M \leftrightarrow W_i \rightarrow Y$ and $M \leftrightarrow W_i \leftrightarrow W_3 \rightarrow Y$;

(ii) $\{W_2, W_1\}$ blocks all backdoor paths from $T$ to $M$ (explicitly: $T \leftrightarrow W_2 \rightarrow M, T \leftrightarrow W_2 \leftrightarrow W_1 \rightarrow M, T \leftrightarrow W_3 \leftrightarrow W_1 \rightarrow M, \text{etc.}$);

(iii) $\{W_3, W_1\}$ blocks all backdoor paths from $\{T, M\}$ to $Y$ (explicitly: $T \leftrightarrow W_3 \rightarrow Y, T \leftrightarrow W_3 \leftrightarrow W_3 \rightarrow Y, T \leftrightarrow W_3 \leftrightarrow W_1 \rightarrow Y, T \leftrightarrow W_2 \leftrightarrow W_2 \leftrightarrow W_3 \rightarrow Y, \ldots$).

We are thus led to the conclusion that the added associations between $W_1, W_2,$ and $W_3$ do not interfere with the identification of $NDE$.

We are also led to appreciate the guidance provided by graphical procedures, without which decisions concerning identification could easily become unmanageable. Fortunately, these procedures are easily mechanizable by present-day software since they are driven entirely by the graph structure. Once a researcher hypothesizes the model structure, a simple algorithm can go through the graphical tests above and, if identifiability is established, deliver the proper mediation formula or estimate it from the data.

The next section discusses examples where the restrictiveness of Assumption Set $B$ may hinder identification and where a careful examination of the $A_{C}$ criteria would be needed to produce unbiased estimates of $NDE$.

### Comparing Identification Power

In comparing the identification power of Assumption Sets $A$ versus $B$, we note that $A$ draws its increased power from two sources:

(a) Divide and conquer—covariates may be found capable of deconfounding the mediator and outcome processes separately but not simultaneously; and

(b) Identification by mediating instruments—intermediate covariates may be measured, enabling one to identify causal effects through multistep procedures, not through a one-step adjustment, as required by $B$.

### Divide and conquer

To highlight the extra power of Assumption Set $A$, we examine the six models in Figure 5. The results of this examination are detailed in Table 2 and can be summarized as follows:

Both $A$ and $B$ deem the $NDE$ identifiable in Models a and $c$ and nonidentifiable in Model d. However, Assumption Set $A$ correctly identifies $NDE$ in Models b, c, and f, while $B$ mistakes it to be nonidentifiable in these models.

The reasoning behind these determinations can best be followed in Figure 5b, which clearly demonstrates how the divide and conquer flexibility translates into increased identification power. Here, there are no backdoor paths from $M$ to $Y$, so $A_{C\rightarrow-2}$ is satisfied by the null set $W = \{\}$.

Still, to deconfound the $T \rightarrow M$ relationship, $A_{C\rightarrow-3}$ requires an adjustment for $W_2$. Likewise, to deconfound the $(T, M) \rightarrow Y$ relationship, $A_{C\rightarrow-4}$ requires an adjustment for $W_3$. If we make the two adjustments separately, both relationships can be deconfounded, and by Corollary 3, $NDE$ reduces to the estimand of Equation 15. However, if we were to adjust for $W_2$ and $W_3$,
simultaneously, as required by Assumption Set B, the $T \rightarrow M$ relationship would become confounded along the path $W_3 \leftarrow Y \leftarrow W_2 \leftarrow M$. In other words, the full set of $A_G$ (or $B_G$) cannot be satisfied by the same set of $W$ elements. As a result, Assumption Set $B$ would deem the $NDE$ to be unidentifiable; there is no covariates set that simultaneously satisfies $B_G^1$ and $B_G^2$.

We note that treatment assignment in this model is not random under any one of the two needed adjustments; $T$ remains confounded (or nonignorable) either with $M$ or with $Y$. It is for this reason that the term deconfounded is less ambiguous than random or as if randomized.

Figure 5f further illustrates why Assumptions A-3 and A-4 insist on identifying $w$-specific effects and, consequently, the extra precautions that this requirement imposes on choosing $W$, even in cases where $NDE$ is identified. If $W = W_1$ is chosen to deconfound the $M \rightarrow Y$ relationship, then $NDE$ can be properly estimated (using $W_2$ to deconfound $T \rightarrow M$ and $W_3$ to deconfound $(T, M) \rightarrow Y)$. However, if $W_3$ is chosen to deconfound the $M \rightarrow Y$ relationship, the $T \rightarrow M$ relationship is no longer deconfoundable, that is, no set of measured variables is available to block all the confounding paths from $T$ to $M$. The conclusion is twofold. First, any software that tells us if $NDE$ is identifiable may need to search the space of candidate sets $W$ before a determination can be made; an independent control for confounding in each of the three relationships, $M \rightarrow Y$, $T \rightarrow M$, and $T \rightarrow Y$, is not sufficient for identifying natural effects. Second, if we venture to skip over this search and estimate the $NDE$ by adjusting for all measured variables, the result is likely to become biased; Figures 5b, 5c, and 5f exemplify this danger.

**Identification by mediating instruments.** Figure 6 displays another model for which Assumption Set $A$ permits the identification of the natural direct effect, while $B$ does not. $NDE$ achieves its identifiability through auxiliary mediating variables ($Z$) but not through adjustment for pretreatment covariates, as demanded by $B$.

In this model, the null set $W = \emptyset$ satisfies Condition B-1 but not Condition B-2; there is no set of covariates that would enable us to deconfound the treatment-mediator relationship. Referring to our encouragement-design example, such a model acknowledges the existence of unmeasured factors that affect both student choice to enroll in the program ($T = 1$) and student ability to devote time for homework ($M = 1$). The intermediate variable, $Z$, that stands between $T$ and $M$ may represent, for example, students’ perception of the importance of homework to their progress, which can be monitored by auxiliary means (e.g., a questionnaire) at some intermediate stage of the study. It can be shown that the availability of such intermediate measurements can make up for the unobservability of all factors that confound $T$ and $M$ (Morgan & Winship, 2007, Chapter 3; Pearl, 2000a, Chapter 3).

Indeed, Condition A-3 requires only that we identify the effect of $T$ on $M$ by some means, not necessarily by rendering $T$ random or unconfounded (or ignorable). The presence of the observed variable $Z$ permits us to identify this causal effect using an estimator called front-door (Pearl, 1995; Pearl, 2009b, pp. 81–85). The resultant $NDE$ estimand will be

$$NDE = \sum_m [E(Y | T = 1, M = m) - E(Y | T = 0, M = m)]$$

$$P(M = m | do(T = 0)), \quad (19)$$

where $P(M = m | do(T = 0))$ is given by

22 This follows from the fact that both $W_1$ and $W_2$ are colliders (i.e., receiving two incoming arrows) along the path; each permits the flow of information when it is conditioned on and stops the flow when not conditioned on (see Appendix A).
The natural direct effect is identified by adjusting for $M$ and not insisting that it be satisfied by the same covariates $Z$. Fortunately, the ability to observe $M$ permits the identification of the effect of $T$ on $Y$. In this model, the mediator $T$ - outcome relationship is unconfounded (since $X$ is a collider), so we are at liberty to choose $W = \{2\}$ to satisfy condition A-2. The treatment $T$ - outcome relationship is confounded and requires an adjustment for $W$. The $\{T, M\} \rightarrow Y$ relationship, however, cannot be deconfounded by any covariate; conditioning on $X$ would confound the $M \rightarrow Y$ relationship, while not conditioning on $X$ would leave the $T \rightarrow Y$ relationship confounded along the path $T \leftarrow X \leftarrow L_1 \rightarrow Y$ (in violation of Condition A-4). Here, the presence of $Z$ comes to our help, for it permits us to estimate $P(y \mid do(t, m))$ using the front-door estimator, as in Equations 19–20, thus rendering NDE identifiable.

### Coping With Treatment-Dependent Confounders

Figure 8 is the first example we encountered in which the natural direct effect is nonidentifiable while the controlled direct effect is identifiable. Another such example is shown in Figure 10 (see Appendix B). Here, $W$ can serve to deconfound both the $M \rightarrow Y$ and the $T \rightarrow M$ relationships, but alas, $W$ is a descendant of $T$, so it violates Condition A-1 and renders NDE nonidentifiable. The controlled direct effect, on the other hand, is easily identifiable using the truncated product formula (see Appendix C). Figure 10 unveils a general pattern that prevents identification of natural effects in any nonparametric model (Avin, Shpitser, & Pearl, 2005; Robins, 2003): Whenever a variable exists, be it measured or unmeasured, that is a descendant of $T$ and an ancestor of both $M$ and $Y$ ($W$ in our examples), NDE is not identifiable.
Figure 10. The natural direct effect is not identifiable because Condition A-1 cannot be satisfied—W is a descendant of T. M = mediator; T = treatment; W = covariate; Y = outcome.

This restriction however does not apply to linear structural models, where parameter identification is all that is needed for the identification of all effects, even when a confounder W of M → Y is affected by the treatment. The reason is that, with the values of all parameters given, the model equations becomes completely specified, from which we can derive all counterfactuals, including those invoked in the definition of natural effect (see Equations 9–10). The same applies to other parametric structural models, such as linear models with interaction terms. This increased identification power comes, of course, at the cost of increasing the danger of misspecification because our commitment to a specific functional form may be incorrect.

To illustrate, consider the parametric version of Figure 10:

\[ y = \beta_1 m + \beta_2 t + \beta_3 tm + \beta_4 w + u, \]
\[ m = \gamma_1 t + \gamma_2 w + u, \]
\[ w = \alpha t + u, \]
with \( \beta_3 m \) representing an interaction term. The basic definition of the natural effects (see Equations 9–10) gives (for the transition from \( T = 0 \) to \( T = 1 \), treating M as the mediator)

\[ NDE(M) = \beta_2 + \alpha \beta_4, \]
\[ NIE(M) = \beta_1 (\gamma_1 + \alpha \gamma_2), \]
\[ TE = \beta_2 + (\gamma_1 + \alpha \gamma_2) (\beta_4 + \beta_1) + \alpha \beta_4, \]
\[ TE - NDE(M) = (\beta_1 + \beta_3) (\gamma_1 + \alpha \gamma_2). \]

We see that, due to treatment-mediator interaction, \( \beta_3 m \), the portion of the effect for which mediation is necessary (\( TE - NDE \)) can differ significantly from the portion from which mediation is sufficient (\( NIE \); Pearl, 2012a). The fact that W is affected by the treatment does not hinder the identification of these effects (as long as the structural parameters are identifiable), though the choice of terms for each of those effects is not trivial and needs to be guided carefully by the formal, counterfactual definitions of \( NDE \) and \( NIE \) (Pearl, 2012b). Even in the simple model of Equations 21–23, with \( \beta_3 \) the only interaction term, it is not at all obvious that \( \beta_3 \) should affect the necessary and sufficient components of mediation in the manner shown in Equations 24–27. The task is much more intricate in the presence of multiple interacting mediators, each acting as both a mediator and a moderator.

For nonparametric models, Avin et al. (2005) derived a necessary and sufficient condition for identifying (natural) path-specific effects in any graph structure with no unmeasured confounders. For example, suppressing the \( T \to W \) or \( T \to M \) processes in Figure 10 would lead to identifiable effects, while suppressing the \( W \to Y \) or \( M \to Y \) processes would not. Shpitser (2013) generalized this result and gave a complete algorithm for path specific effects with multiple treatments, multiple outcomes, and hidden variables.

Figure 10 can in fact be regarded as having two interacting mediators, M and W, and the results of Avin et al. (2005) highlight a fundamental difference between the two. Whereas effects mediated through W are identifiable, those mediated through M are not. For example, the natural direct and indirect effects viewing the mediator can be obtained directly from Equations 16 and 17, exchanging \( m \) with \( w \), since the relationships \( T \to W \) and \( T W \to Y \) are unconfounded. This gives

\[ NDE(W) = \sum_w \{ E(Y \mid T = 1, W = w) - E(Y \mid T = 0, W = w) \} P(W = w \mid T = 0), \]
\[ NIE(W) = \sum_w E(Y \mid T = 0, W = w) [P(W = w \mid T = 1) - P(W = w \mid T = 0)], \]

in which M is not invoked, since it is regarded as part of the direct effect from T and Y.24

For comparison, the parametric version of Figure 10 given in Equations 21–23 yields the following effects when M is considered the mediator:

\[ NDE(W) = \beta_2 + \gamma_1 \beta_1. \]
\[ NIE(W) = \alpha (\beta_4 + \gamma_2 \beta_1). \]
\[ TE = \beta_2 + (\gamma_1 + \alpha \gamma_2) (\beta_4 + \beta_1) + \alpha \beta_4, \]
\[ TE - NDE(W) = \alpha (\gamma_2 \beta_3 + \beta_4 + \gamma_2 \beta_1 + \gamma_1 \beta_3). \]

Comparing Equations 28–31 to Equations 24–27 allows an investigator to assess the relative contribution of each mediator, W and M, to the overall effect of T on Y.

Figure 11 depicts the parameterized model of Equations 21–23 and compares the subgraphs carrying the effects (NIE) mediated by M and W, respectively.

Conclusions

I have presented a concise, general, and interpretable set of conditions for identifying natural effects, and demonstrated by examples how they can be tested in a given model and how they lead to improved identification power. In particular, the new conditions open the door for identification methods that go beyond standard adjustment for covariates and leverage auxiliary variables and multistep procedures that operate in the presence of confounded treatment and mediator relationships.

Footnotes:

23 Such models have been analyzed extensively in the literature, some using a purely statistical approach (Jo, 2008; Kraemer et al., 2008; Mackinnon, 2008; Preacher, Rucker, & Hayes, 2007) and some applying the mediation formula of Equations 16 and 17 (Coffman & Zhong, 2012; Imai, Keele, & Yamamoto, 2010; Muthén, 2011; Pearl, 2010a, 2012a; Valeri & VanderWeele, 2013; VanderWeele & Vansteelandt, 2009). However, the problem of dealing with two interacting mediators (e.g., M and W in Figure 10) has not received much attention.

24 Remarkably, if W were merely correlated with M, rather than causally affecting it, the effect mediated by either M or W would not be identified, since no measured covariate can satisfy Assumptions A-1 and A-2.
Applying these conditions to linear models with interaction terms, I have shown how path-specific effects can be estimated in models with multiple pathways and interacting mediators.

An important feature of the conditions formulated in this article is their *mechanizability*. Simple graphical algorithms exist (and are cited in the reference list) that examine the structure of the model, test whether the identification conditions are satisfied in the model, and, depending on how they are satisfied, produce an unbiased estimate of the desired mediated effect. This feature relieves researchers from the task of interpreting and judging the validity of each identifying assumption in isolation; it is the plausibility of the postulated model structure (i.e., the diagram) that one needs to judge and defend. The structure itself dictates both the choices by which the identification conditions can be satisfied and the estimation procedures appropriate for each choice.

Naturally, to apply these identification procedures to real-life data, one needs to be certain of the causal scenario behind the data and that the scientific context of that scenario is faithfully depicted in the diagram. The question arises whether it is realistic to assume that investigators would possess such certainties in real-life applications. Here, one should recall that anchoring one’s analysis in specific causal scenarios does not imply a commitment to the validity of those scenarios. It implies willingness to explore their ramifications, to evoke critiques of one’s assumptions, and to understand which variants of each scenario are critical for identification and for choosing the correct estimator. The alternative, of course, is to sweep these uncertainties under the rug of no unmeasured confounders or sequential ignorability. This article replaces such sweeping assumptions with specific scientific contexts (encoded graphically) that investigators can scrutinize for plausibility, submit to statistical tests, and appeal to mechanical procedures for identification analysis. This departure from ignorability-based approaches to mediation should provide researchers with a deeper understanding of the nature of mediation and the tools available for its analysis.

25 The testable implications of causal diagrams are discussed in Appendix A (see Bollen & Pearl, 2013; Pearl, 2009b, pp. 140–144).

References


Pearl, J. (2012a). The causal mediation formula: A guide to the assessment of


(Appendices follow)
Appendix A

Covariate Selection: d-Separation and the Backdoor Criterion

Consider an observational study where we wish to find the effect of treatment (T) on outcome (Y), and assume that the factors deemed relevant to the problem are structured as in Figure A1; some are affecting the outcome, some are affecting the treatment, and some are affecting both treatment and response. Some of these factors may be unmeasurable, such as genetic trait or lifestyle, while others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment so that if we compare treated versus untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of factors is called a sufficient set, admissible set, or a set appropriate for adjustment (see Greenland, Pearl, & Robins, 1999; Pearl, 2000b, 2009a). In this article, I call such a set a deconfounder of the T → Y relationship.

I now describe a criterion named backdoor (Pearl, 1993), which provides a graphical method of selecting such a set of factors for adjustment. It is based on the simple idea that, when we adjust for a set S of covariates, we should block, or disable, all spurious paths from T to Y and leave intact all causal paths between the two. To operationalize this idea, we need the notion of d-separation (the d stands for directional), which provides a formal characterization of what it means to block a path and also allows us to detect all the testable implications that a given model entails.

Definition 1 (d-separation): A set S of nodes is said to block a path p if either (a) p contains at least one arrow-emitting node that is in S or (b) p contains at least one collision node that is outside S and has no descendant in S. If S blocks all paths from set T to set Y, it is said to d-separate T and Y, and then, variables T and Y are independent given S, written $T \perp \!\!\!\perp Y|S$.\(^\text{A1}\)

The intuition behind d-separation can best be recognized if we regard paths in the graph as conveyers of probabilistic information, with nodes acting as information switches. In causal chains $i \rightarrow m \rightarrow j$, the two extreme variables are marginally dependent but become independent of each other (i.e., blocked) once we condition on (i.e., know the value of) the middle variable. Figuratively, conditioning on m appears to block the flow of information along the path, since learning about i has no effect on the probability of j, given m. Inverted forks $i \leftarrow m \rightarrow j$, representing two causes having a common effect, act the opposite way; if the two extreme variables are (marginally) independent, they become dependent (i.e., connected through unblocked path) once we condition on the middle variable (i.e., the common effect) or any of its descendants. This special handling of collision nodes (or colliders), 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.

To illustrate, the path $Z_1 \rightarrow W_1 \rightarrow T$ in Figure A1 is blocked by $S = W_1$, and the path $Z_1 \rightarrow Z_2 \rightarrow T$ is blocked by $S = Z_3$, since each of these nodes emits an arrow along its corresponding path. Moreover, all other paths from Z to T (e.g., $Z_1 \rightarrow Z_3 \rightarrow Y \leftarrow W_1 \leftarrow T$) are blocked by $S = \{Z_2\}$, since Y is a collider. Consequently, the set $S = \{W_1, Z_3\}$ d-separates $Z_3$ from $T$, and we can conclude that the conditional independence $Z_1 \perp T|\{W_1, Z_3\}$ will be satisfied in any probability function that this model can generate, regardless of how we parameterize the arrows.

Similarly, the path $Z_1 \rightarrow Z_3 \leftarrow Z_2$ is blocked by the null set $\emptyset$, but it is not blocked by $S = \{Y\}$ since Y is a descendant of the collision node $Z_3$. Consequently, the marginal independence $Z_1 \perp Z_2$ will hold in the distribution, but $Z_1 \perp Z_2|Y$ will most likely not hold.

Each conditional independence implied by a d-separation condition in the diagram offers a statistical test that can be performed on the data to confirm or refute the validity of the model. These tests can easily be enumerated by attending to each missing edge in the graph and selecting a set of variables that d-separate the pair of variables corresponding to that missing edge.

For example, in Figure A1, three of the missing edges are $Z_2 \rightarrow Z_1$, $Z_1 \rightarrow Y$, and $Z_2 \rightarrow T$, with separating sets $\emptyset$, $\{T, Z_2, Z_3\}$, and $\{Z_1, Z_3\}$, respectively. Accordingly, the testable implications of $M$ include $Z_2 \perp Z_1, Z_1 \perp Y|\{T, Z_2, Z_3\}$, and $Z_2 \perp T|\{Z_1, Z_3\}$.

In linear systems, these conditional independence constraints translate into zero partial correlations, or zero coefficients in the corresponding regression equations. For example, the three implications above translate into the following constraints: $r_{Z_1Z_2} = 0, r_{YZ_1} \neq r_{Z_2Z_1} = 0$, and $r_{Z_1T}Z_2Z_3 = 0$.

\[\text{A1}\] In other words, the conditional independence $T \perp Y|S$ can be shown to hold in every distribution that the model can generate, regardless of the functional form of the equations in the model and regardless of the distribution of the omitted factors (Pearl & Verma, 1991). See Hayduk et al. (2003), Mulaik (2009), Elwert (2013), and Pearl (2009b, p. 335) for gentle introductions to d-separation.
Such tests are easily conducted by routine regression techniques, and they provide valuable diagnostic information for model modification, in case any of them fails (see Pearl, 2009b, pp. 143–145). Software routines for automatic detection of all such tests, as well as other implications of graphical models, are reported in Kyono (2010).

Armed with the tool of $d$-separation or path blocking, we are ready to tackle the issue of identification using the backdoor criterion. This criterion provides a graphical method of selecting admissible sets of factors and demonstrates that causal quantities such as $P(y \mid do(t))$ can often be identified with no knowledge of the functional form of the equations or the distributions of the latent variables in $M$.

**Definition 2 (admissible sets—the backdoor criterion):** A set $S$ is admissible (or sufficient) for estimating the causal effect of $T$ on $Y$ if two conditions hold:

1. No element of $S$ is a descendant of $T$.
2. The elements of $S$ block all backdoor paths from $T$ to $Y$—namely, all paths that end with an arrow pointing to $T$.

Based on this criterion we see, for example, that, in Figure A1, the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, $\{W_1, Z_3\}$, and $\{W_2, Z_3\}$ (among others) are each sufficient for adjustment because each blocks all backdoor paths between $T$ and $Y$. The set $\{Z_1\}$, however, is not sufficient for adjustment because it does not block the path $T \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

The intuition behind the backdoor criterion is as follows. The backdoor paths in the diagram carry spurious associations from $T$ to $Y$, while the paths directed along the arrows from $T$ to $Y$ carry causative associations. Blocking the former paths (by conditioning on $S$) ensures that the measured association between $T$ and $Y$ is purely causal, namely, it correctly represents the target quantity: the causal effect of $T$ on $Y$. The reason for excluding descendants of $T$ (e.g., $W_i$ or any of its descendants) are discussed in Appendix C, while conditions for relaxing this restriction are given in Pearl (2009b, p. 338) and Shpitser, VanderWeele, and Robins (2010). The implication of finding a sufficient set, $S$, is that stratifying on $S$ is guaranteed to remove all confounding bias relative to the causal effect of $T$ on $Y$. In other words, it renders the causal effect of $T$ on $Y$ identifiable, via

$$P(Y = y \mid do(T = t)) = \sum_s P(Y = y \mid T = t, S = s)P(S = s).$$  \hfill (A1)

Since all factors on the right-hand side of the equation are estimable (e.g., by regression) from preinterventional data, the causal effect can likewise be estimated from such data without bias. Moreover, the counterfactual implication of $S$ can be written as $T \perp Y \mid S$, also known as conditional ignobility (Rosenbaum & Rubin, 1983).

The backdoor criterion allows us to write Equation A1 by inspection, after selecting a sufficient set, $S$, from the diagram. The selection criterion can be applied systematically to diagrams of any size and shape, thus freeing analysts from judging whether “$T$ is conditionally ignorable given $S$,” a formidable mental task required in the potential-response framework. The criterion also enables the analyst to search for an optimal set of covariates—namely, a set, $S$, that minimizes measurement cost or sampling variability (Tian, Paz, & Pearl, 1998).

(Appendices continue)
Appendix B

Formal Derivation of Conditions for Natural Direct Effect Identification (After Pearl, 2001)

Notation

We retain the notation used in the rest of the article and let $T$ be the control variable (whose effect we seek to assess) and $Y$ be the response variable. We let $M$ stand for the set of all intermediate variables between $T$ and $Y$ that, in the simplest case considered, would be a single variable $M$ as in Figure 4 in the main text.

We use the counterfactual notation $Y(u)$ to denote the value that $Y$ would attain in unit (or situation) $U = u$ under the control regime $do(T = t)$. See Equation 4 in the main text and Pearl (2000a, Chapter 7) for formal semantics of these counterfactual expressions. Many concepts associated with direct and indirect effect require comparison to a reference value of $T$, that is, a value relative to which we measure changes. We designate this reference value by $t^*$. Natural Direct Effects: Formulation

Definition 3 (unit-level natural direct effect; qualitative):

An event $T = t$ is said to have a natural direct effect on variable $Y$ in situation $U = u$ if the following inequality holds:

$$ Y_t(u) \neq Y_{t^*,u}(u). \quad \text{(B1)} $$

In words, the value of $Y$ under $T = t$ differs from its value under $T = t^*$ even when we keep $M$ at the same value ($M_{t^*}(u)$) that $M$ attains under $T = t^*$. We can easily extend this definition from events to variables by defining $T$ as having a natural direct effect on $Y$ (in model $M$ and situation $U = u$) if there exist two values, $t^*$ and $t$, that satisfy Equation B1. Note that this definition does not require that we specify a value $m$ for $M$; that value is determined naturally by the model, once we specify $t, t^*$, and $u$.

If one is interested in the magnitude of the natural direct effect, one can take the difference

$$ Y_{t,M_{t^*},u}(u) - Y_{t^*,u}(u) \quad \text{(B2)} $$

and designate it by the symbol $NDE(t, t^*; Y, u)$ (acronym for natural direct effect). If we are further interested in assessing the average of this difference in a population of units, we have the following:

Definition 4 (average natural direct effect): The average natural direct effect of event $T = t$ on a response variable $Y$, denoted $NDE(t, t^*; Y)$, is defined as

$$ NDE(t, t^*; Y) = E(Y_{t,M_{t^*}}) - E(Y_t). \quad \text{(B3)} $$

Natural Direct Effects: Identification

As noted in Robins and Greenland (1992), we cannot generally evaluate the average natural direct effect from empirical data. Formally, this means that Equation B3 is not reducible to expressions of the form

$$ P(Y = y) \text{ or } P(Y_{m} = y); $$

the former governs the causal effect of $T$ on $Y$ (obtained by randomizing $T$), and the latter governs the causal effect of $T$ and $M$ on $Y$ (obtained by randomizing both $T$ and $M$).

We now present conditions under which such reduction is nevertheless feasible.

Theorem 2 (experimental identification): If there exists a set $W$ of covariates, nondescendants of $T$ or $M$, such that

$$ Y_{t,m} \perp M_{t^*} \mid W \quad \text{for all } m \quad \text{(B4)} $$

(read: $Y_{t,m}$ is conditionally independent of $M_{t^*}$, given $W$), then the average natural direct effect is experimentally identifiable, and it is given by

$$ NDE(t, t^*; Y) = \sum_{w,m} \left[ E(Y_{t,m} \mid w) - E(Y_{t^*,m} \mid w) \right] \frac{P(M_{t^*} = m \mid w)P(w)}{P(M_{t^*} = m \mid W = w)P(W = w)}. \quad \text{(B5)} $$

Proof

The first term in Equation B3 can be written

$$ E(Y_{t,M_{t^*}} = y) = \sum_{w,m} \sum_{w} E(Y_{t,m} = y \mid M_{t^*} = m, W = w) \frac{P(M_{t^*} = m \mid W = w)P(W = w)}{P(M_{t^*} = m \mid W = w)P(W = w)}. \quad \text{(B6)} $$

Using Equation B4, we obtain

$$ E(Y_{t,M_{t^*}} = y) = \sum_{w} \sum_{w} E(Y_{t,m} = y \mid W = w) \frac{P(M_{t^*} = m \mid W = w)P(W = w)}{P(M_{t^*} = m \mid W = w)P(W = w)}. \quad \text{(B7)} $$

Each factor in Equation B7 is identifiable: $E(Y_{t,m} = y \mid W = w)$, by randomizing $T$ and $M$ for each value of $W$, and $P(M_{t^*} = m \mid W = w)$, by randomizing $T$ for each value of $W$. This proves the assertion in the theorem. Substituting Equation B7 into Equation B3 and using the law of composition $E(Y_{t^*}) = E(Y_{t^*,M_{t^*}})$ (Pearl 2000a, p. 229) gives Equation B5 and completes the proof of Theorem 2.

(Appendices continue)
The conditional independence relation in Equation B4 can easily be verified from the causal graph associated with the model. Using a graphical interpretation of counterfactuals (Pearl, 2000a, pp. 214–215), this relation reads

\[(Y \perp M \mid W)_{G_{TM}} \]  \hspace{1cm} (B8)

In words, \( W \) \( d \)-separates \( Y \) from \( M \) in the graph formed by deleting all (solid) arrows emanating from \( T \) and \( M \).

Figure B1a illustrates a typical graph associated with estimating the direct effect of \( T \) on \( Y \). The identifying subgraph is shown in Figure B1b and illustrates how \( W \) separates \( Y \) from \( M \). The separation condition in Equation B8 is somewhat stronger than Equation B4, since the former implies the latter for every pair of values, \( t \) and \( t' \), of \( T \) (see Pearl 2000a, p. 214).

The identification of the natural direct effect from nonexperimental data requires stronger conditions. From Equation B5, we see that it is sufficient to identify the conditional probabilities of two counterfactuals: \( P(Y_{tm} = y \mid W = w) \) and \( P(M_{t'} = m \mid W = w) \), where \( W \) is any set of covariates that satisfies Equation B4 (or Equation B8). This yields the following criterion for identification:

**Theorem 3 (nonexperimental identification):** The average natural direct effect \( \text{NDE}(t, t'; Y) \) is identifiable in nonexperimental studies if there exists a set \( W \) of covariates, nondescendants of \( T \) or \( M \), such that for all values \( m \) and \( w \) we have

(i) \( Y_{tm} \perp M_{t'} \mid W; \)

(ii) \( P(Y_{tm} = y \mid W = w) \) and \( P(Y_{tm'} = y \mid W = w) \) are identifiable; and

(iii) \( P(M_{t'} = m \mid W = w) \) is identifiable.

Moreover, if Conditions (i)–(iii) are satisfied, the natural direct effect is given by Equation B5, in which all counterfactual expressions are replaced by their probabilistic estimands.

In particular, for confounding-free models, we obtain the mediation formulas of Equations 16–17 in the main text.

(Appendices continue)
Appendix C

Why Treatment-Dependent Covariates Cannot Be Used to Deconfound the Mediator-Outcome Process

Assumption Sets A and B both insist that no member of W be affected by the treatment, which is a requirement distinct to the identification of natural effects. For example, to identify the controlled direct effect CDE(m) in Figure 10 in the main text, we can condition on \( W = w \), and using the truncated product formula (Pearl, 2000a, p. 72), we can write

\[
CDE(m) = E[Y \mid do(T = 1, M = m)] - E[Y \mid do(T = 0, M = m)]
= \sum_w E[Y \mid T = 1, M = m, W = w]P(T = 1, W = w) - E[Y \mid T = 0, M = m, W = w]P(T = 0, W = w).
\]

The reason such conditioning does not work for the natural direct effect is that the latter is defined not in terms of a population operation (i.e., control M to level \( M = m \), and change \( T \) from \( T = 0 \) to \( T = 1 \)) but in terms of a hypothetical manipulation at the unit level, namely, for each individual \( u \), freeze \( M \) at whatever level it attained for that individual, then change \( T \) from \( T = 0 \) to \( T = 1 \) and observe the change in \( Y \).

Appendix A shows that in order to convert this unit-based operation to a population-based operation (expressible as a do(t) expression), we must first find a \( W \) that deconfounds \( M \) from \( Y \) (with \( T \) fixed) and then, conditioned on that same \( W \), identify the counterfactual expression

\[
P(M_i = m \mid W = w).
\]

When \( W \) is affected by the treatment, this expression is not identifiable even when \( T \) is randomized. To see that, we recall that \( M_i \) stands for all factors affecting \( M \) when \( T \) is held fixed. These factors are none other but the omitted factors (or disturbance terms) that affect \( M \), namely, \( U_{iM} \) in Figure 1 in the main text. When we condition on \( W \), those factors become correlated with \( T \), which renders \( T \) confounded with \( M \).

This can also be seen from the graph, using virtual colliders. The expression \( P(M_i = m \mid W = w) \) stands for the causal effect of \( T \) on \( M \) within a stratum \( w \) of \( W \). It is identifiable using the backdoor criterion, which demands that \( W \) not be affected by \( T \) because, as soon as \( W \) is a descendant of any intermediate variable from \( T \) to \( M \) (including \( M \) itself), a virtual collider is formed and a new backdoor path is opened by conditioning on \( W \) (Pearl, 2009b, p. 339).

Another way of seeing this is to resort to do-calculus. If \( W \) is not affected by the treatment, we have \( W_i = W \), and we can write

\[
P(M_i = m \mid W = w) = P(M_i = m \mid W_i = w)
= \frac{P(M_i = m, W_i = w)}{P(W_i = w)}
= \frac{P(M_i = m, W = w \mid do(T = t))}{P(W = w \mid do(T = t))}
= P(M = m \mid do(T = t), W = w).
\]

The last expression stands for the causal effect of \( T \) on \( M \) given that \( W = w \) is the posttreatment value of \( W \). It is identifiable by the do-calculus, whenever the model permits such identification (Shpitser & Pearl, 2008).

(Appendices continue)
It is worth mentioning at this point that treatment-dependent confounders hinder only nonparametric identification of natural effects as defined in Equation B3. The difficulty disappears when we have a parametric representations (as in Equations 21–23 in the main text) or when we compromise on the requirement of freezing \( M \) completely at the value it attained prior to the change in treatment. For example, if, in Figure 10 in the main text, we merely disable the process \( T \rightarrow M \) and allow \( M \) to respond to \( W \) as we change \( T \) from \( T = 0 \) to \( T = 1 \), the resulting direct effect will be identified. These types of direct and indirect effects, which I would like to call seminatural effects,\(^1\) are defined (using parenthetical notation) as

\[
\begin{align*}
SNDE &= E[Y(T = 1), M(T = 0, W(T = 1)), W(T = 1)] - E[Y(T = 0)], \\
SNIE &= E[Y(T = 0), M(T = 1, W(T = 0)), W(T = 0)] - E[Y(T = 0)].
\end{align*}
\]

Using the derivation leading to Equation B5, one can show that these seminatural effects are identifiable by

\[
\begin{align*}
SNDE &= \sum_{m,w} E(Y \mid T = 1, M = m, W = w)P(M = m \mid T = 0, W = w) \\
&\quad \cdot P(W = w \mid T = 1) - E(Y \mid T = 0), \\
SNIE &= \sum_{m,w} E(Y \mid T = 0, M = m, W = w)P(M = m \mid T = 1, W = w) \\
&\quad \cdot P(W = w \mid T = 0) - E(Y \mid T = 0).
\end{align*}
\]

Accordingly, the parametric model of Equations 21–23 in the main text would yield the following seminatural effects:

\[
\begin{align*}
SNDE &= \beta_2 + \alpha(\beta_4 + \gamma_2\beta_1), \\
SNIE &= \gamma_1\beta_1, \\
TE &= \beta_2 + (\gamma_1 + \alpha\gamma_2)(\beta_1 + \beta_2) + \alpha\beta_4, \\
TE - SNDE &= \gamma_1(\beta_1 + \beta_2) + \beta_2\alpha\gamma_2.
\end{align*}
\]

\(^1\) Huber (2012) called them partial indirect effects.

Figure C1. Subgraphs supporting the seminatural direct and indirect effects (\( SNDE \) in panel a, \( SNIE \) in Panel b) and those supporting the natural direct and indirect effects (\( NDE \) in Panel c, \( NIE \) in Panel d). \( M \) = mediator; \( T \) = treatment; \( W \) = covariate; \( Y \) = outcome; \( \alpha, \beta, \gamma \) = structural coefficients.

Figure C1 depicts the path that supports the \( SNDE \) (seminatural direct effect) and \( SNIE \) (seminatural indirect effect) compared with those supporting the \( NDE \) (natural direct effect) and \( NIE \) (natural indirect effect) in Equations 24–27 in the main text. We see that the criterion of Avin et al. (2005) is satisfied in the latter, but not the former.