Electronic Supplementary Material to Accompany

Causal Effects and Counterfactual Conditionals: Contrasting Rubin, Lewis and Pearl

Keith A. Markus

John Jay College of Criminal Justice of The City University of New York

This document provides a summary of notation, brief introductions to Rubin's causal model (RCM), Lewis's theory of counterfactual conditionals (LTC), and Pearl's structural causal model (SCM), and some further details on the relationships between the formal systems representing LTC and SCM. The intention of the summaries is to provide sufficient background to make the main text of the article accessible to readers unfamiliar with one or more of the three theories.

Notation

The following notation is used throughout.

Symbol	*Meaning*
~	Negation: logical not (~P same as $\bar{P}$)
∧	Conjunction: logical 'and'
∨	Disjunction: logical 'or' (weak or, permitting both disjuncts to be true)
⊃	Material conditional: truth-functional if-then (P ⊃ Q same as Q ∨ ~P, same as ==>)
≜	Definitional equivalence
□→	Lewis conditional: Counterfactual conditional according to LTC. (Differs from Stalnaker conditional.)
ω	Denotes a possible word
Do(x), Do(X = x)	Pearl's Do() operator. Result of manipulating model to delete equation for *X* and replace *X* with *x*.
$Y_x(u)$	Holland's notation for RCM potential response
□	Necessity operator
◊	Possibility operator

Rubin's Causal Model

Rubin's Causal Model offers a means of conceptualizing causal effects as a basis for estimation of those effects (Imbens & Rubin, 2015; Rubin, 1974). The model begins with the assumption that causal effects inherently involve comparisons between at least two different treatments. Treatments differ both with respect to the manner in which one brings them about and the further consequences that these differences bring. Unpacking this notion of causal effect thus begins with two key components: first, the specification of the treatment interventions that give rise to them and, second, the conceptualization and formal representation of the outcomes to which the treatments give rise.

Assume a population of individual units, indexed by *i* ranging over natural numbers. For simplicity, focus on a single treatment contrast between two treatments. Treatments constitute inherently compound events comprising both (a) the differences between the treatments to be compared and (b) the common prior conditions of the units before the intervention that distinguishes the two treatments. Specifying the treatment conditions to be compared determines the structure of counterfactuals to be considered later because in RCM everything pivots on the treatments being studied.

Once one has specified the treatments, $W = 0$ and $W = 1$, it becomes possible to posit outcomes in the assigned treatments for each unit. $Y_i(w)$ equals the outcome on variable Y for unit i in condition $W = w$. RCM labels $Y_i(0)$ and $Y_i(1)$ potential outcomes. The individual causal effect for unit i equals $Y_i(0) - Y_i(1)$. One can think of RCM in terms of possible worlds logic in which treatment assignment varies across worlds and thus so do outcomes for those treatments for various units. The treatment assignment mechanism plays a central role in RCM with respect to explicating estimation of causal effects. However, Rubin conceptualized potential outcomes as (relational) properties of units in the actual world.

The potential outcomes notation introduced above tacitly assumes that for any $Y_i(w) = y$, y exists and has a determinate value. This receives explication in the Stable Unit Treatment Value Assumption (SUTVA, or Stability Assumption for short) which has two parts, the second of which does not receive consistent recognition in the literature. The first part states that the potential outcomes for unit *i* do not depend upon the treatment assignments for units other than *i*. The second part states that potential outcomes for unit *i* do not vary within the assigned treatment. The latter part might be violated, for example, if there were inconsistency in the dosage of aspirin between pills. This second part is sometimes expressed as a prohibition against hidden variables.

The *Fundamental Problem of Causal Inference* involves the fact that the researcher can only observe one outcome for each unit (Rubin, 1978). Thus, one can conceptualize the problem of causal inference as a missing data problem (Imbens & Rubin, 2015). Imagine a complete data set that included a set of covariates, X, the treatment assignment, W, and the potential outcomes $Y_i(0)$ and $Y_i(1)$. For the purposes of RCM, units are not repeatable. The same individual at two different times counts as two different units. As a consequence, each unit can receive only one treatment. As such, each unit will only have an observed outcome for one treatment condition and the rest will remain missing. The observed outcome will equal whichever potential outcome corresponds to the assigned treatment for that unit. Note that the observed outcome will exactly equal that potential outcome because all factors, systematic or random, that impact the outcome are baked into the potential outcome. The conceptualization of potential outcomes in RCM is thus closely tied to the idea that the causal inference problem is a missing data problem.

One can represent the *assignment mechanism* as some function $\text{pr}(W \mid X, Y(0), Y(1))$ such that (a) it makes no difference in what order one lists the units and (b) the probabilities across treatment conditions sum to 1. Rubin and colleagues have considered various limiting assumptions that can

facilitate causal inference in various ways. *Strongly Ignorable Treatment Assignment* combines three such restrictions (Imbens & Rubin, 2015; Rosenbaum & Rubin, 1983). First, each unit's probabilities of assignment are not unduly influenced by assignments or outcomes for other units. Second, each unit has a non-zero probability of assignment to each treatment. Third, assignment probability does not depend upon the potential outcomes conditional on X: $\text{pr}(W \mid X, Y(0), Y(1)) = \text{pr}(W \mid X)$. It warrants noting that this last condition does not imply a lack of association between W and Y because Y takes its value from different potential outcomes depending upon W.

As an illustrative example, RCM helps us think through the estimation of the effect of using a coaster. If one randomly assigns tables (units) to coaster use treatments (W), then water rings (Y) will reflect the two sets of potential outcomes $Y_i(W = \text{coaster})$ and $Y_i(W = \text{no coaster})$. Random assignment to W renders the potential outcomes independent of W. If instead, W depended upon table type (X), then table type might induce an association between treatment condition and potential outcomes. However, conditioning on table type would restore conditional independence, again allowing for an unbiased estimate. Rubin's methodology often focuses on reducing a multivariate set of covariates to a single *propensity score* which contains all the information about the probability of treatment contained in the covariates. A fundamental feature of RCM is the effort to use methodology to separate considerations related to unbiased estimation, represented by the propensity score model, from considerations related to the phenomenon under study, what Rubin (2004a, 2004b) calls the science, represented by the model of the outcome variable. The goal is to facilitate researchers making decisions about method in a way that is not biased by hypotheses about the science.

Lewis' Theory of Counterfactual Conditionals and Comparative Possibility

Natural language supports a variety of conditional assertions. The material conditional of first-order logic, $A \supset B$, probably offers the best known example. The semantics of the material conditional

provide rules to determine its truth value, *true* or *false*, based on the truth values of the propositions represented by the logical variables *A* and *B*. The material conditional is defined to hold true except when *A* holds true and *B* holds false (in the context of a classical bivalient logic where every proposition is determinately either true or false). As such, it always holds true when *A* holds false. However, natural language also supports conditional statements that can hold true or false even if they have a false antecedent (in the above example, *A* is the antecedent and *B* the consequent of the conditional). In some contexts it makes sense to speak of what would have happened had some antecedent been true with the understanding that assertions that some things would have happened might come out false. For example, let *A* = "The economy adds enough jobs" and *B* = "Labor force participation reaches 100%", assuming both individually false. It seems false that if *A* then *B* so long as factors such as wages produce voluntary unemployment. Such reasoning requires a different kind of conditional from the material conditional, commonly referred to as a counterfactual conditional. Lewis (1973b, 1973c) provided a theory of such counterfactual conditionals.

The material conditional fails to account for such conditionals because it takes its truth value entirely from the actual state of affairs whereas counterfactual conditionals depend for their truth values on other possible states of affairs. Similarly to Stalnaker (1968), Lewis represented these states of affairs as worlds, with the actual world representing the actual state of affairs within which the truth value is sought. The *possibility* of proposition *A* in the actual world corresponds to the accessibility to the actual world of at least one world in which *A* holds true. From any one possible world, other possible worlds can be either accessible or inaccessible. Moreover, Lewis assumed some weak ordering of possible worlds in terms of comparative similarity such that the ordering meets two conditions. The first condition requires a *connected* and *transitive* ordering. This means that for any two possible worlds, one falls as close or closer to the actual world than does the other in terms of

similarity. Further, if ω_1 falls at least as close as ω_2 ($\omega_1 \leq \omega_2$) and $\omega_2 \leq \omega_3$ then $\omega_1 \leq \omega_3$. Second, Lewis made a *centering assumption* that the actual world remains accessible to itself and closer to itself than any other world. With these two assumptions, Lewis offered the following analysis of counterfactual conditionals. "$A \,\square\!\!\rightarrow C$ is true at i iff some (accessible) AC-world is closer to i than any $A\,\bar{C}$ -world if there are any (accessible) A-worlds" (Lewis 1973b, p. 424-425). Lewis denoted the actual world as i, worlds where A and C hold true as AC worlds and worlds where A holds true but not C as $A\,\bar{C}$ worlds. Lewis used '$\square\!\!\rightarrow$' to denote the counterfactual conditional connective and distinguish it from the material conditional. For a variety of reasons, Lewis understood possible worlds as members of a multiverse no different in kind from the actual world aside from the fact that we inhabit the actual world rather than some other world (Lewis, 1986a).

To illustrate, consider the counterfactual conditional asserting "If I had not used a coaster then my glass would have left a ring on the table." Assume that in the actual world I used a coaster and left no ring. Further assume that my drink can leave rings and that my table can sustain them.
Does the assertion hold true? If one focuses only on the two propositions joined within the conditional, it may initially seem as though a world in which I do not use a coaster but leave no ring appears more similar to a world in which I do use a coaster and do not leave a ring. However, Lewis considers the entire world up until the moment of difference when evaluating similarity. A non-coaster non-ring world would differ in terms of true generalities about coaster use and ring leaving. Considered this way, we then judge the non-coaster ring worlds more similar than the non-coaster non-ring worlds. In that case, the conditional comes out true. It bears noting, however, that evaluated at some more distant world in which the table is made from a different substance or my drink is served at a different temperature, the same conditional might come out false.

Also note that Lewis does not appeal to causation in his theory of counterfactual conditionals. He sometimes refers to laws but these are universally quantified descriptive statements, not causal laws. The avoidance of causation has a purpose. Lewis (1973a, 1986b, 2004) proposed a counterfactual theory of causation that endeavored to explicate the concept of causation in terms of counterfactual conditionals. Thus, had his theory of counterfactual conditionals relied on any causal concepts, it would have risked circularity, failing in its effort to explicate causation using only non-causal concepts.

Structural Causal Model

Pearl (2009) offered an account of (some) causal counterfactual conditionals based on structural causal models (SCMs; the term applies to both individual models and the theory that characterizes them). A SCM comprises three sets: A set, U, of exogenous variables, a set, V, of endogenous variables,[1] and a set, F, of functions from the values of U and V to the values of V such that the value of each V_n in V depends only upon the corresponding U_n in U and a subset of V excluding V_n[2]. Drawing from graph theoretic terminology, the members of V on which a variable depends in its function in F constitute the *parents* of the variable and the converse constitute the *children* of the variable. If one

1 There exists some ambiguity as to whether the distinction between endogenous and exogenous variables involves *model-implied endogeneity* (common in the structural equation modeling literature) or what one might call *real-world endogeneity* independent of the model determined by how the world works (common in the econometrics literature). The ambiguity arises because Pearl (2009, section 7.1) appears to have assumed a correct model that corresponds to how the world works. However, the presentation appears entirely formal and Pearl never discussed a variable being wrongly identified as exogenous or endogenous in the subsequent text. So, it remains a plausible conjecture that the most felicitous interpretation of Pearl's exposition implies that one should call a U variable exogenous in a model even if the variable so represented does not exhibit exogeneity outside the model. Nothing important will turn on this issue in what follows.

2 The restriction that only one U appear in each equation excludes most latent variable models. However, one can often relax this restriction without harm, as in the M-bias example. The requirement also conflicts with common terminology in structural equation modeling in which exogenous variables can be observed variables and an equation can contain more than one of them (Bollen, 1989). Nothing of importance will turn on this terminological issue.

applies these relations recursively, one obtains the *ancestors* and *descendants* of the variable. The SCM approach assumes a unique solution for the values of all the variables conditional on values for the variables in *U* at least for recursive models.

Pearl introduced a Do(...) operator to clearly distinguish probabilistic statements from causal statements. Suppose one were to pile half one's coasters face up in a left pile, and the other half face down in a right pile. *pr*(*Facing* = *up* | *Pile* = *left*) = 1 because probabilities describe the actual state of affairs. Nonetheless, moving a downward facing coaster into the left pile will not make it face upwards. pr(*Facing* = *up* | (Do[*Pile* = *left*] & *Facing* = *down*)) = 0, because moving coasters between piles does not flip them over. The implicit grammar of the such statements interprets the expression before the pipe, |, as referring to the time after the intervention whereas conditions that follow the pipe refer to pre-intervention values and restrict the selection of cases to which the intervention applies. To determine this value, one deletes the function for *Pile* from the model and replaces it with *Pile* = *left*. One then uses the remaining functions to solve for the new values of the other variables. The function for *Facing* does not depend upon *Pile* and so the downward facing coaster selected by the condition in the Do(...) statement retains its value. Pearl borrowed the term 'potential outcome' defined in his system as the value of $Y \mid \mathrm{Do}(X = x)$ written $Y_{X=x}(u)$ for unit u[3]. Thus a causal counterfactual conditional asserting that if $X = x$ then $Y = y$ holds true just in case $[Y \mid \mathrm{Do}(X = x)] = y$.

SCM defines causal effects in terms of the function from one variable to the probability distribution of another, $P(Y) = f(x)$, rather than just a difference in expected value: $E(Y_{X=1}(u)) - E(Y_{X=0}(u))$. Stochastic models over populations of units combine a structural causal model with a probability distribution over the values of *U*. Pearl adopted the idiom of referring to unbiased estimates of causal effects as identified[4]. Identification of a causal effect in a parametric model means not only

3 The use of 'u' for units reflects the interchangeability of units with the same values for the *U* variables in SCM.

4 This usage contrasts with common use in structural equation modeling where a biased estimator with a determinate value counts as identified whereas in this usage the quantity of interest is considered not identified even if a determinate

that one can compute a unique value from the combination of a SCM and passively observed data on the variables included in the model, but that the resulting estimate varies around the population value of the effect. Pearl provides an *Adjustment Formula* for estimating such effects as well as various graphical criteria for evaluating identification. Unlike RCM, SCM makes integral use of directed graphs and Pearl encouraged the use of these in applied problem solving. Pearl (e.g., 2012) also considers non-parametric models for which identification applies to specific questions bearing determinate answers.

The graphical criteria generally revolve around a key concept termed *d-separation*. A *path* from node X to node Y of a graph refers to any sequence of edges connecting nodes of the graph leading from X to Y such that each edge ends on the same node at which the edge that follows it begins, irrespective of the direction of any of the edges. A set of nodes, $\boldsymbol{Z}$, d-separates X and Y along a path if it meets at least one of the following two criteria. (Note that in this notation capital bold letters denote sets and capital non-bold letters denote elements of a set. This has been regularized from Pearl's notation in which non-bold capital letters indicate sets and lower case letters indicate elements in order to clearly distinguish it from the case in which non-bold capital letters represent scalar variables and lower case letters represent determinate values of a variable.) First, the path contains a chain from L to M to N or a fork from M to both L and N such that node M is a member of set $\boldsymbol{Z}$. Second, the path contains an inverted fork from L and N to M such that M is not in $\boldsymbol{Z}$ and no descendant of M is in $\boldsymbol{Z}$. The set $\boldsymbol{Z}$ d-separates X and Y if and only if the nodes in $\boldsymbol{Z}$ d-separate X and Y along every path between the two. The set $\boldsymbol{Z}$ d-separates the set $\boldsymbol{X}$ and the set $\boldsymbol{Y}$ if and only if $\boldsymbol{Z}$ d-separates each node in $\boldsymbol{X}$ from each node in $\boldsymbol{Y}$. The *Back-door Criterion* for identification of the effect of X on Y requires that a

---

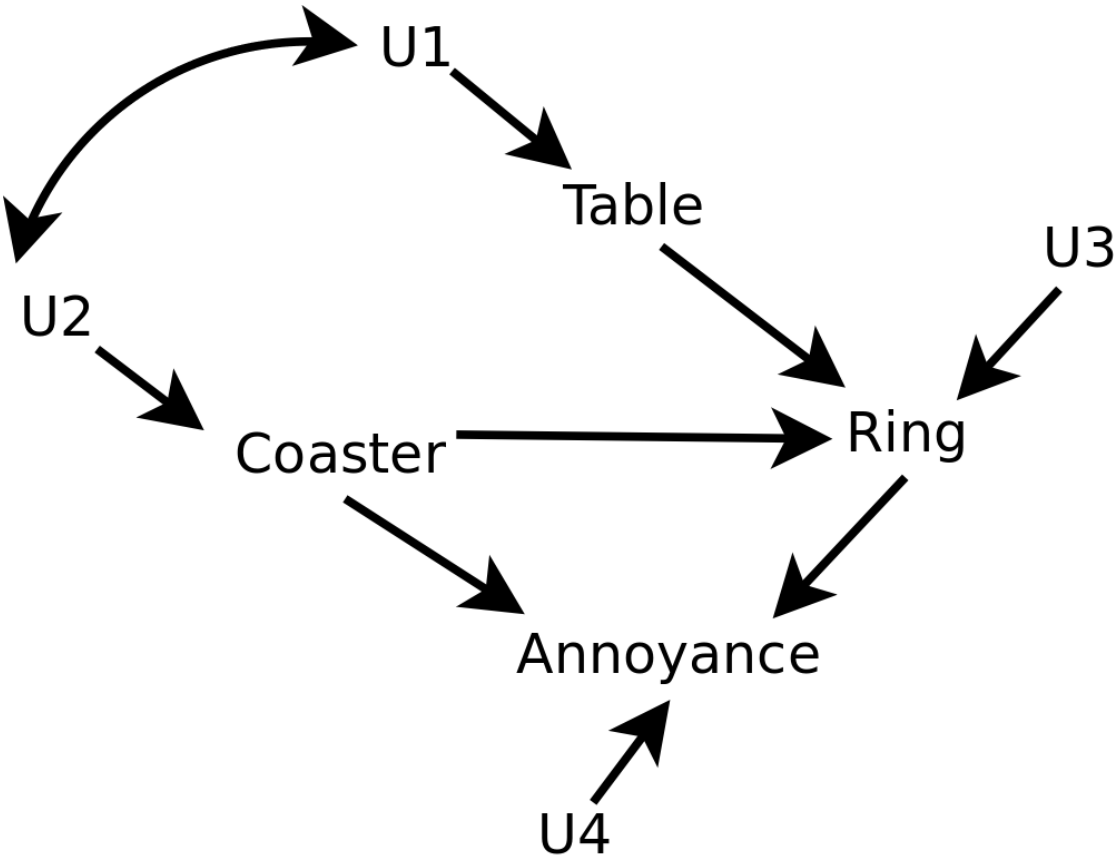
estimated value exists (Bollen, 1989).

sufficient set of covariates d-separates *X* from *Y* for all paths that begin with a directed edge pointing toward *X* without incorporating any descendants of *X*.

Also unlike RCM in which one builds the model around the causal effect one wishes to estimate, SCM treats all the variables in the model democratically such that one could use the same model to reason about more than one causal effect. Figure 1 presents a simple directed graph incorporating the coaster example (notice that directed graphs adopt a more abstract representation than path diagrams and therefore do not correspond one-to-one to statistical models the way that path diagrams do). The association between coaster use and water rings fails to provide an unbiased estimate of the corresponding causal effect but one can obtain an unbiased estimate by controlling for table type. In contrast, controlling for host annoyance would introduce bias into the estimate of the causal effect. The table node d-separates the coaster node and the ring node. The set containing the table node and the annoyance node fails to d-separate the other two nodes.

Figure 1

Illustrative Causal Graph



Further Details Comparing the Formalisms of LTC and SCM

The relation of formal isomorphism discussed in the main text offers a necessary but insufficient condition for strong equivalence. It is necessary because if two theories do not share the same formal structure, then they cannot have the same content. It is insufficient because two isomorphic theories can nonetheless diverge in content due to differences between the two theories in the meanings of the formal symbols. This section adds a few additional comments on the lack of formal isomorphism between LTC and SCM with implications for strong equivalence.

Conditionals Involving Quantification over Cases

Pearl (2009: Ch. 1) introduced the general form of nonparametric structural equations in terms of scalar values of variables, implying that the scalar functions are applied individually to different

cases, or at least applied to vectors such that each value of the variable on the left-hand side is determined only by the corresponding values of the variables on the right-hand side. That is, one person's value on a causal variable does not determine someone else's value on the effect variable (similar to the first clause of SUTVA in RCM). Consider a counterfactual conditional with a universally quantified consequent, quantified over cases: e.g., "If one stock plunges, they all will." There is no problem evaluating such a conditional using LTC. Evaluating such a conditional using SCM requires some form of extension or precisification of the existing canonical exposition. Do we allow vector functions such that a value for one case depends upon a value for another? Do we simply prohibit such functions as beyond the domain of causal relations? If so, do we evaluate them all as false or do we decline to assign any truth value? Alternatively, do we adopt a middle strategy that allows some but not others (e.g., "If any one stock price is above average then they all are.")? SCM offers no ready-made answer, the answer remains indeterminate. Nonetheless, there is no reason to doubt that one or more answers may be workable. Such matters offer fruitful avenues for the further development of SCM (e.g., Halpern, 2000; Briggs, 2012; Fisher, 2017), but offer little probative value because there exists no clear demarcation between developing strategies implicit in SCM and developing strategies that extend SCM to handle such cases.

Conditional Law of Excluded Middle

The Conditional Law of Excluded Middle states $(A \mathrel{\square\!\!\rightarrow} B) \vee (A \mathrel{\square\!\!\rightarrow} \sim B)$ which one can rewrite as $\sim(\sim(A \mathrel{\square\!\!\rightarrow} B) \wedge \sim(A \mathrel{\square\!\!\rightarrow} \sim B))$ bringing into relief Lewis's intuition that both conditionals could hold false in a given situation. Given that B and $\sim B$ exhaust the possibility space for the consequent, this follows from the axioms in Pearl's system. The key axiom states that $\exists x \in X$ such that $X_y(u) = x$, which Galles & Pearl, 1998, called *Definiteness* and Pearl, 2009, relabeled *Existence*. The first instance of X presumably denotes the range of X and the range of B is T and F, thus either $B_a(u) = T$ or

$B_a(u) = F$, which is exactly what the Conditional Law of Excluded Middle requires. The *Uniqueness* axiom requires that $\sim((A \mathrel{\Box\!\!\rightarrow} B) \wedge (A \mathrel{\Box\!\!\rightarrow} \sim B))$ in this example, which brings out a related difference because if one substitutes a contradiction for A then LTC yields the result that both conditionals hold true.

Modal Expressions

Aside from the fact that counterfactual conditionals of the form $A \mathrel{\Box\!\!\rightarrow} \Diamond A$ violate SCM's assumption that all counterfactual conditionals rest on causal determination, the lack of any means of handling modal propositions in SCM produces other problems as well. Separate nodes for A and $\Diamond A$ would either violate modularity or allow (A and $\sim\Diamond A$) which implies a different accessibility relation among possible worlds than that assumed by LTC. Moreover, if a possible world in SCM is a set of value assignments to U, then there is no clear mechanism for modeling semantic dependencies across possible worlds. None of these difficulties arise for LTC for which modal antecedents and modal consequents pose no special challenges. LTC does not assume a causal similarity metric and does not restrict antecedents or consequents to extensional statements about properties borne by property bearers. $\Diamond A$ holds true in ω if and only if A holds true in some world accessible to ω and $\Box A$ (necessarily A) holds true in ω if and only if A holds true in every world accessible to ω (Hughes & Cresswell, 1996).

Nested Conditionals and Import/Export

Accepting Import/Export involves the rejection of Modus Ponens (McGee,1985). Intuitions that favor the idea that antecedents express irreversible interventions incline in the direction of giving up Modus Ponens for counterfactual conditionals. Intuitions that favor the idea of reversible antecedents instead reject Import/Export in favor of Modus Ponens. See Briggs (2012) and Fisher (2017) for further discussion.

References

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies, 160,* 139-166. doi: 10.1007/s 11098-012-9908-5

Fisher, T. (2017). Causal counterfactuals are not interventionist counterfactuals. *Synthese, 194,* 4935-4957. doi: 10.1007/s11229-016-1183-0

Galles, D. & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science, 3,* 151-182. doi: 10.1023/A:1009602825894

Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research, 12,* 317-337. doi: 10.1613/jair.648 (Previously published 1998)

Hughes, G. E. & Cresswell, M. J. (1996). *A new introduction to modal logic.* London: Routledge.

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction.* New York: Cambridge University Press.

Lewis, D. (1973a). Causation. *Journal of Philosophy, 70,* 556-567. doi: 10.2307/2025310

Lewis, D. (1973b). *Counterfactuals*. Malden, MA: Blackwell.

Lewis, D. (1973c). Counterfactuals and comparative possibility. *Journal of Philosophical Logic, 2,* 418-446.

Lewis, D. (1986a). *On the plurality of worlds*. Malden, MA: Blackwell.

Lewis, D. (1986b). *Philosophical papers* (vol. 2). Oxford, UK: Oxford University Press.

Lewis, D. (2004). Causation as influence. In J. Collins, N. Hall & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 75-106). Cambridge, MA: MIT Press.

McGee, V. (1985). A counterexample to modus ponens. *The Journal of Philosophy, 82,* 462-471.

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

Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle, Ed., *Handbook of structural equation modeling* (pp. 68-91). New York: Guilford.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology, 66,* 688-701. doi: 10.1037/h0037350

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6,* 34-58. doi: 10.1214/aos/1176344064.

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

Rubin, D. B. (2004b). Reply to discussion. *Scandinavian Journal of Statistics, 31,* 196–198.

Stalnaker, R. (1968). A Theory of Conditionals. In N. Rescher (Ed.), *Studies in Logical Theory* (pp. 98-112). Oxford, UK: Blackwell.