Two Theorems on Invariance and Causality*

Nancy Cartwright[†][‡]

In much recent work, *invariance under intervention* has become a hallmark of the correctness of a causal-law claim. Despite its importance this thesis generally is either simply assumed or is supported by very general arguments with heavy reliance on examples, and crucial notions involved are characterized only loosely. Yet for both philosophical analysis and practicing science, it is important to get clear about whether invariance under intervention is or is not necessary or sufficient for which kinds of causal claims. Furthermore, we need to know what counts as an intervention and what invariance is. In this paper I offer explicit definitions of two different kinds for the notions *intervention, invariance,* and *causal correctness.* Then, given some natural and relatively uncontroversial assumptions, I prove two distinct sets of theorems showing that invariance is indeed a mark of causality when the concepts are appropriately interpreted.

1. Introduction.

1.1. The project. Much recent work on causal inference takes *invariance under intervention* as a mark of correctness in a causal-law claim (Glymour, Scheines, Spirtes, and Kelly 1987; Hausman and Woodward 1999; Hoover 2001; Redhead 1987). Often this thesis is simply assumed; when it is argued for, generally the arguments are of a broad philosophical nature with heavy reliance on examples. Also, the notions involved are

*Received March 2000; revised November 2001

[†] To contact the author, please write to Nancy Cartwright, Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, England, or Philosophy Department, 0119, UCSD, 9500 Gilman Drive, La Jolla, CA 92093–0119.

[‡]Thanks to Daniel Hausman and James Woodward for setting me off on this project and two referees for helpful suggestions. This research was funded by a grant from the Latsis Foundation, for which I am grateful, and it was conducted in conjunction with the Measurement in Physics and Economics Project at LSE. I wish to thank the members of that group for their help, especially Sang Wook Yi and Roman Frigg.

Philosophy of Science, 70 (January 2003) pp. 203–224. 0031-8248/2003/6904-0013$10.00

https://doi.org/10.1086/367876 Published online by Cambridge University Press

often characterized only loosely, or very specific formulations are assumed for the purposes of a particular investigation without attention to a more general definition, or different senses are mixed together as if it did not matter. But it does matter, because a number of different senses appear in the literature for each of the concepts involved, and the thesis is false if the concepts are lined up in the wrong way.

To get clear about whether invariance under intervention is or is not necessary or sufficient for a causal-law claim to be correct, and under what conditions, we need to know what counts as an intervention, what invariance is, and what it is for a causal-law claim to be correct. Next we should like some arguments that establish clear results one way or the other. In this paper I offer explicit definitions for two different versions of each of the three central notions: *intervention, invariance,* and *causal claim.* All of these different senses are common in the literature. Then, given some natural and relatively uncontroversial assumptions, I prove two distinct sets of theorems showing that invariance is a mark of causality when the concepts are appropriately interpreted. These, though, are just a sample of results that should be considered.

The two different sets of theorems use different senses of each of the three concepts involved and hence make different claims. Both might loosely be rendered as the thesis that a certain kind of true relation will be invariant when interventions occur. In the second, however, what counts as "invariance" becomes so stretched that the term no longer seems a natural one, despite the fact that this is how it is sometimes discussed in the literature—especially by James Woodward, whose extensive study of invariance is chiefly responsible for isolating this particular characteristic and focussing our attention on it.

Nor is "intervention" a particularly good label either. The literature on causation and invariance is often connected with the move to place manipulation at the heart of our concept of causation (Price 1991; Hausman 1998; Woodward 1997; Hausman and Woodward 1999): roughly, part of what it means to be a cause is that manipulating a cause is a good way to produce changes in its effects. "Manipulation" here I take it suggests setting the target feature where we *wish* it to be, or *at will,* or *arbitrarily.* Often when authors talk about intervention, it sounds as if they assume just this aspect of manipulation.

Neither set of theorems requires a notion so strong. All that is required is that nature allow specific kinds of *variation* in the features under study.[1] We might argue that manipulability of the right sort will go a good way towards ensuring the requisite kind of variability. But mere variation of

[1] Or, if the right kind of variation does not actually occur, there must be a fact of the matter about what would happen were it to do so.

https://doi.org/10.1086/367876 Published online by Cambridge University Press

the right kind will be sufficient as well, so we need take care that formulations employing the terms “manipulation” and “intervention” not mislead us into demanding stronger tests for causality than are needed.

In this paper I am concerned only with claims about deterministic systems where the underlying causal laws are given by linear equations linking the size of the effect with the sizes of the causes. Although this is extremely restrictive, it is not an unusual restriction in the literature, and it will be good to have some clean results for this well-known case. The next step is to do the same with different invariance and intervention concepts geared to more general kinds of causal systems and less restrictive kinds of causal-law claims.

This project is important to practicing science. When we know necessary or sufficient conditions for a causal-law claim to be correct, we can put them to use to devise real tests for scientific hypotheses. And here we cannot afford to be sloppy. Different kinds of intervention and invariance lead to different kinds of tests, and different kinds of causal claims license different things we can do. So getting the definitions and the results straight matters to what we can do in the world and how reliable our efforts will be.

1.2. The Nature of Deterministic Causal Systems. I need in what follows to distinguish between causal laws and our representations of them; I shall use the term “causal system” for the former, “causal structure” for the latter. I take it that the notion of a “causal law” cannot be reduced to any non-modal notions. So I start from the assumption that there is a difference between functional relations that are just true and ones that are true in a special way; the latter are nature’s causal laws. I will also assume transitivity of causal laws. This implies that the causal systems under study include not only facts about what causal laws are true—*e.g.,* “Q causes P”—but also about the possible *ways* by which one factor can cause another—*e.g.,* “Q causes P via R and S but not via T.”

I discuss only linear systems, and I shall represent nature’s causal equations like this: $q_e c = \Sigma a_{ej} q_j$, with the effect on the left and causes on the right. As will be clear from axiom A_1, this law implies that $q_e = \Sigma a_{ej} q_j$; but not the reverse. Following the distinction between systems and structures, I shall throughout use q_i to stand for quantities in nature and x_i for the variables used to represent them. Also with respect to notation, I shall use lower case letters for variables and quantities and upper case letters for their values. I assume the following about nature’s causal systems:

A_1: *Functional dependence.* Any causal equation presents a true functional relation.

A_2: *Antisymmetry and irreflexivity.* If q causes r, r does not cause q.

A_3: *Uniqueness of coefficients.* No effect has more than one expansion in the same set of causes.

A_4: *Numerical transitivity.* Causally correct equations remain causally correct if we substitute for any right-hand-side factor any function in its causes that is among nature's causal laws.

A_5: *Consistency.* Any two causally correct equations for the same effect can be brought into the same form by substituting for right-hand-side factors in them functions of the causes of those factors given in nature's causal laws.

A_6: *Generalized Reichenbach principle.* No quantities are functionally related unless the relation follows from nature's causal laws.

More formally: a *linear deterministic system (LDS)* is an ordered pair $<Q, CL>$, where the first member of the pair is an ordered set of quantities $<q_1, \ldots, q_m>$ and the second is a set of causal laws of the form q_k c$=\Sigma_{j<k} a_{kj} q_j$ (a_{kj} a real number) that satisfies A_1 through A_6.[2]

2. Causal Law Variation, Invariance, and One Kind of Causal Claim.

2.1. The First Definitions. The kind of intervention we shall be concerned with in this section is the same as employed by Pearl (2000b) in his work on causal counterfactuals and by Glymour, Scheines, Spirtes, and

[2] More precisely, a *causal law* for an effect x_j, $L(x_j)$, is a set of ordered pairs giving causes of x_j and their weights: $L(x_j) = \{<a(1)_{j1},x_1><a(2)_{j1},x_1> \ldots <a(k_1)_{j1},x_1><a(1)_{j2},x_2> \ldots <a(k_n)_{jn},x_n>\}$, $a(k)_{jm} \varepsilon$ R. We can then define *x_i causes x_j with weight a* just in case $\exists L(x_j)$ $(<a,x_i> \varepsilon L(x_j))$. (Notice that my formulations allows—as I have argued we should—for a cause to have multiple capacities with respect to the same effect. Once we have admitted this piece of information we can of course go on to define some concept of "the overall influence" of a given cause on a given effect).

Clearly the assumptions too need a more precise formulation. *Transitivity,* for example, becomes

A′4: For any laws $L(x_j)$ and $L(x_i)$, and for any $<b,x_i> \varepsilon L(x_j)$, $L'(x_j)$ is also a law, where

$$\begin{aligned} L'(x_j) = L(x_j) - \{<b,x_i>\} \cup \\ \{<b,a'(1)_{il}, x_i>, \ldots, <b,a'(k_1)_{in}, x_i>\} \\ \text{for all} <a',(k_m)_{im}, x_m> \varepsilon L(x_i) \end{aligned}$$

The other assumptions are formulated similarly.

We need some kind of complicated formulation like this to make clear, e.g., that arbitrary regroupings on the right-hand side of the causal-law equation will not result in a causal law. For example, assume that x_2 c $= ax_1$ and x_3 c $= bx_1 + cx_2$. It follows that $x_3 = bx_1 + (c\text{-}d)x_2 + dx_2 = bx_1 + (c\text{-}d)ax_1 + dx_2 = (b + ca\text{-}da)x_1 + dx_2$, but we do not wish to allow that x_3 c $= (b + ca\text{-}da)\ x_1 + dx_2$. For our purpose here, I think we can proceed with the more intuitive formulations in the text.

Kelly (1987) in their manipulation theorem (once we transform their notion from graph representations to linear deterministic systems). It is also one of the kinds that Daniel Hausman and James Woodward (1999) discuss in their joint work on the Markov condition.

As I indicated in Section 1.1, the results I aim to establish are not really results about *intervention* in the natural sense of that term, but rather results about *variation.* The first kind of intervention, which will be under discussion here in Section 2, is one in which causal laws vary; in the second kind, which I discuss in Section 3, it is the values of the causes picked out in a fixed causal system that vary. We may perhaps be more used to thinking of quantities as taking on different values than of laws as varying.[3] But all we need here is that there are different causal systems that relate to each other in the specific way I shall describe. The point I am trying to make is that it is the occurrence of these systems[4] that matters for testing the correctness of causal claims; it is not necessary that they come to occur through anything naturally labeled an *intervention* or a *manipulation.*[5] I shall, therefore, talk not of *intervention* but rather, of *variation.*

In the first kind of "variation"/"intervention," which I call *causal-law variation,* a new causal system is considered, similar in many ways to the first. Let us call the new system a *test system* for results of quantity q relative to the original system. The *test system* differs from the original that we wish to test by exactly one addition and two kinds of deletions. For a target quantity q, add the law $q = Q$ for some specific value, Q, of q within its allowed range. Drop (1) all laws with q as effect and (2) all laws linking causes of q with effects, e, of q where the causal influence passes through q—that is, any equation for e that can be obtained by transitivity from an equation giving q's effects on e. The first is easy to say formally: drop all laws of the form q c = f(. . .). The second is more cumbersome: drop any equation A: e c = f(. . . , g(. . .), . . .) for which there are equations of the form B: e c = f(. . . , q, . . .) and C: q c = g(. . .).

As with "intervention," there are a number of different kinds of invariance suggested in the literature. The one relevant here seems genuinely a

[3] In my own work (Cartwright 1999) on laws it is natural that they should vary since laws are epiphenomena, depending upon stable arrangements of capacities. I take the prevalence of "intervention" tests for causal correctness of the kind described here, based on the possibility of variations in causal laws, to indicate that a surprising number of other philosophers are committed to something like my view.

[4] Or, the possibility of the occurrence of these systems. (See footnote 1.)

[5] There are of course other kinds of arguments for linking manipulation and causation (e.g., Hausman 1998, Price 1991). My point here is that it is mistaken to argue that manipulation is central to causation *on the grounds that* one important kind of test for causal correctness—the "invariance" test—cannot do without it.

https://doi.org/10.1086/367876 Published online by Cambridge University Press

notion of *invariance,* so that is what I shall call it. An equation in a (linear deterministic) causal system <Q, CL> giving a true functional relation (but not necessarily one that replicates one of nature's causal laws) is *invariant* in q iff it continues to give a true functional relation for any value that q takes in any test situation for q relative to <Q, CL>.

We also need to be explicit about what an equation of the form $x_e c = \Sigma a_i x_i$ in a causal representation is supposed to be claiming. I propose the obvious answer: an equation of this form claims to record one of nature's causal laws. When it does so, I shall say that it is causally correct.

2.2. The First Theorem.

Theorem 1. A functionally true equation is causally correct iff it is invariant in all its independent variables, either singly or in any combination.

Correctness → Invariance

The result in this direction is trivial now that the background is in place. Consider an equation that is causally correct:

$$\text{E: } x_e c = f(x_1, \ldots, x_n).$$

Consider a test system for the effects of q_i for any q_i represented by an x_i in the right-hand side of E. The intervention replaces the causal system of which this equation is a part by a new one. This equation would be dropped from the new system if it had q_i as an effect—which it hasn't. Otherwise it would be dropped only if it had as effect an effect of q_i—which it has—and it results from substituting g(. . .) for q_i into some equation for q_e, where $q_i\, c = g(. . .)$. But in this case q_i would no longer appear in the equation to be dropped. So $x_e\, c = f(x_1, . . . , x_i, . . . , x_n)$ will still obtain in the new system. Hence E is invariant under interventions on q_i.

Clearly the trick in establishing the necessity of invariance for correctness is in the characterization of interventions. So we shall need to be wary when we introduce a different concept of intervention, as in Section 3.

Invariance → Correctness

Consider an equation

$$\text{F: } x_e = \sum_{i=1}^{N} a_i x_i$$

where either some x_i appears that it is not the cause of x_e, or, if all are genuine causes, some x_i appears with a causally incorrect coefficient. In

https://doi.org/10.1086/367876 Published online by Cambridge University Press

order to be invariant, F must also be derivable in all test systems for all q_i and it must be derivable from the same equations as in the original. That is because the move to a test system adds only one kind of new law to use in a derivation: "$q_i = Q_i$" where Q_i may be any value in the appropriate range. This clearly will not help since Q_i will vary from test system to test system, and F must be derivable in all of them. But if F is derivable from the same set of laws in the test situation as in the original, then not only will F be invariant in all x_i, so too must each member of this set be. So we wish to establish:

Lemma 1

> No matter what the causal system, no linear combination of nature's causal equations will yield an equation of form F that is invariant in all the q_i represented on the right-hand side of F.

We should first notice that, trivially,

> *Claim 1.* No matter what the causal system, no causal equations used in the linear combination can have an x_i on the left-hand side.

The result is then established by coupling Claim 1 with

> *Claim 2.* No matter what the causal system, no linear combination of causal equations in which x_i's appear only on the right-hand side will yield F.

Proof of Claim 2. The proof of Claim 2 is by induction on the number of variables in addition to x_e and the x_i's that appear in the equations in the linear combination that yields F.

Inductive Base. As a base for the induction, show that no linear combination of equations in any causal system that use *no* variables in addition to x_e and the x_i's and are invariant in all x_i will yield F. Here's how: All equations used in such a linear combination will have x_e on the left-hand side and some combination of x_i's on the right-hand side. That is, they will look like this:

$$\mathrm{B}: x_e\ c = \sum b_i x_i$$

$$\mathrm{C}: x_e\ c = \sum c_i x_i$$

$$\vdots$$

where some of the b_i and some of the c_i will be zero. By *consistency,* some combination of factors from B cause factors in C or the reverse or both. But if factors in B cause a factor represented by[6] x_i in C, then B will not

[6] I shall henceforth drop the use of "represented by" where it will not cause confusion and simply talk of variables causing other variables.

https://doi.org/10.1086/367876 Published online by Cambridge University Press

be invariant in x_i. Similarly, if factors in C cause a factor, x_i', in B, then C will not be invariant in x_i'. So no two such equations can be used and F cannot be so obtained.

InductiveArgument. We aim to establish by reductio that if Claim 2 is true for a set of equations using n variables in addition to x_e and the x_i's, it will be true for a set using $n+1$ additional variables. So suppose F can be obtained using $n+1$ additional variables; let $z_1, \ldots, z_k$, $k = N + n + 1$, denote the variables that appear in a linear combination that yields F.

> *Lemma 2.* At least one of the "extra variables"—one of the z_i that is neither x_e nor any of the x_i's—must appear as an effect in the equations used at least once. Call it z.
>
> *Proof.* This is true because
>
> (i) Among extra variables that appear as causes, at least one will not be a cause of any of the other extra variables involved. Otherwise we would have a causal loop, which violates *antisymmetry.* Call it z'.
>
> (ii) Since z' does not appear in F, it must appear in at least two equations (one to introduce it, one to eliminate it).
>
> (iii) Both these equations must have x_e as effect since no x_i can appear as an effect in an invariant equation. z' could appear with the same coefficient in both equations:

$$x_e = az' + \Sigma a_i z_i$$

$$x_e = az' + \Sigma b_j z_j$$

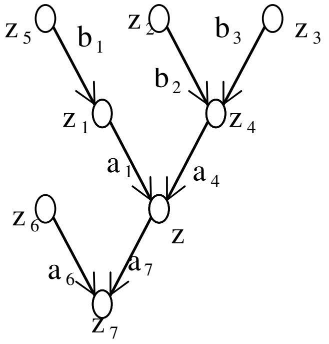
By *consistency,* $\Sigma\, a_i z_i$ and $\Sigma\, b_j z_j$ can be brought into the same form by a set of laws, *L,* linking the z_i and the z_j. In this case these two equations containing z' can be replaced in F by the laws in *L,* which do not contain z', with no loss. Alternatively, z' can appear with different coefficients in the two equations:

$$x_e = az' + \Sigma a_i z_i$$

$$x_e = bz' + \Sigma b_j z_j$$

But this is possible only if z' is a cause of either one or more of the z_i or of the z_j. Since these effects must be x_i's, the equation with the causes of these x_i's will not be invariant in all x_i.

https://doi.org/10.1086/367876 Published online by Cambridge University Press



$$a_2 =_{df} a_4 b_2$$
$$a_3 =_{df} a_4 b_3$$
$$a_5 =_{df} a_1 b_1$$

We can now eliminate z in the following way: consider nature's causal law for z as effect that cites as causes just those factors that are direct causes of z among the z_i. Designate it thus:

$$z \text{ c} = \Sigma a_i y_i \quad y_i \in \{z_i, \ldots, z_k\}$$

Replace any equation in the original linear combination in which z appears as cause by the same equation with $\Sigma a_i y_i$ substituted for z. Eliminate all equations with z as effect. Add nature's causal equations giving the relations among all the causes that appear in all the different equations that had z as effect, as well as those connecting z′s parents with the effects of z among the z_i. For example, supposing the relations in Figure 1, we replace

$$z \text{ c} = a_1 z_1 + a_2 z_2 + a_3 z_3$$

$$z \text{ c} = a_4 z_4 + a_5 z_5$$

$$z_7 \text{ c} = a_6 z_6 + a_7 z$$

with

$$z_1 \text{ c} = b_1 z_5$$

$$z_4 \text{ c} = b_2 z_2 + b_3 z_3$$

$$z_7 \text{ c} = a_6 z_6 + a_7 (a_1 z_1 + a_4 z_4)$$

Clearly the new set of equations will be invariant in all x_i if the original are, and any equation in x_e and the x_i that can be obtained using the original equations can be obtained using the new ones. Q.E.D.

3. Variation of Values, Prediction of First Differences, and Parameter Correctness.

3.1. Systems That Are Nice for Us. The basic idea in connecting intervention/variation with invariance as a test of causality is Mill's method of concomitant variation: as a cause changes, the effect should change "in train." But there are caveats. The variation must occur in the right circumstances. The easiest circumstances are where the putative cause varies all on its own and no other causes vary at all. That is essentially what we achieve in the test systems of Section 2 by looking at variants of the original causal laws that make the putative cause take a particular value independent of what values other factors have.

But sometimes, if a causal system is sufficiently nice, we can achieve essentially the same results by looking within the system itself. The simplest case is where each of the putative causes for a given effect has a cause of its own that can vary without any cross restraints on other possible causes of that effect. That will guarantee that all possible causes can take on any combination of values. I call such a system *epistemically convenient.*

More formally, an *epistemically convenient linear deterministic system (ECLDS)* is a linear deterministic system, <Q, CL>, such that

> A_7: *Epistemological convenience.* For each q_j in $Q = \{q_1, \ldots, q_m\}$ there is some cause q_j^* such that:
>
> (i) $q_j\ c = \Sigma_{k<j}c_{jk}q_k + q_j^*$
> (ii) There are no cross restraints on the values of the q_j^*; that is, for all situations in which <Q, CL> obtains, it is possible ("allowed by nature") for each q_j^* to take any value in its allowed range consistent with all other q_k^* taking any values in their allowed ranges.[7]

In case the LDS we are studying is an epistemically convenient one, we can relabel the quantities so that the system takes the familiar form

$$q_1\ c = u_1$$

[7] This is similar to a standard kind of condition on parameter values in econometrics (cf. Engle, Hendry, and Richard 1983) and as a condition on parameter values plays a central role in Kevin Hoover's (2001) theory of causal inference. Woodward (1997) asks for statistical independence of the exogenous quantities. The proof here requires the additional assumption that there are no cross restraints on their values.

$$q_2 \ c = a_{21}q_1 + u_2$$

$$\vdots$$

$$q_n \ c = a_{n1}q_1 + \ldots + a_{nn-1}q_{n-1} + u_n,$$

where $n = m/2$. For the remainder of this part, I consider only epistemically convenient linear deterministic systems, and I assume that the notation has its natural interpretation for such systems.

Notice that (i) and (ii) imply

(iii) no q_k in Q causes q_j^*

but neither

(iv) for all j, k, q_j^* does not cause q_k^*

nor

(v) for all j, k, q_j^* and q_k^* have no common cause (i.e., they are not part of any other LDS in which they have a common cause).

Many authors restrict their attention to systems satisfying (iv) and (v) as well, usually with the intention of mounting an argument from (i), (iii), (iv), and (v) to (ii). I shall not do so because the argument is not straightforward and at any rate we need only the assumption (ii) for deriving the results of interest here.

Following standard usage, let us call the "special causes" represented by *u*'s in an ECLDS, *exogenous* quantities, since they are not caused by any quantities in the system. Notice that, for an ECLDS, an assignment of values to each of the exogenous quantities will fix the values of all other quantities in the system. In what follows it will help to have an expression for a quantity in the system in terms of the exogenous quantities. Again following conventional usage, I call this the *reduced form.*

$$\text{RF: } q_k \ c = \sum_{i=1}^{k} u_i \sum_{l=i}^{k-1} a_{kl} \sum_{m=i}^{l-1} a_{lm} \ldots$$

where we adopt the convention $\sum_{j=\alpha}^{\beta} f_i(j,k,l,\ldots) = 1$, if $\alpha > \beta$.

$$\therefore q_k \ c = \sum_{i=1}^{k} \Gamma_i^k u_i,$$

where $\Gamma_i^k = \sum_{l=i}^{k-1} a_{kl} \sum_{m=i}^{l-1} a_{lm} \ldots.$

Call any set of values for each of the exogenous terms a situation. We shall be interested in differences so let us define $\Delta_j^\alpha q_n = \text{df } q_n(u_1 = U_1, \ldots, u_{j-1} = U_{j-1}, u_j = U_j + \alpha, u_{j+1} = U_{j+1}, \ldots, u_{m/2} = U_{m/2})$—$q_n(u_1 = U_1, \ldots, u_{j-1} = U_{j-1}, u_j = U_j, u_{j+1} = U_{j+1}, \ldots, u_{m/2} = U_{m/2})$.

Statisticians like epistemologically convenient systems because they make estimation of probabilities from data easier. We, by contrast, are concerned with how to infer causal claims given facts about association. For this project, these kinds of systems have three advantages.

1) In Section 2 we discussed methods for finding out about a causal system of interest by looking at what happens in *other* related systems. But the existence of the system of interest provides no guarantee that these other systems exist for us to observe. In this part we shall be interested in situations in which specified factors take arbitrary values relative to each other. In an epistemologically convenient system this is guaranteed to happen "naturally" within the system itself—at least "in the long run."[8]

2) Consider a *functionally correct* hypothesis,

$$H: x_e \ c = \Sigma a_{ej} x_j$$

where each q_j (represented by x_j) has an exogenous cause peculiar to it satisfying ii). In this case nature provides a basic arrangement that allows the possibility for each q_j to have an *open back path;* whether indeed each does have an open back path will depend entirely on our knowledge, but at least the facts are right to allow us knowledge of the right kind. Relative to q_e, q_j has an open back path just in case (a) every causal law with q_j as effect has a u_j such that u_j cannot cause q_e except by causing q_j, and (b) we know what these u's are and we know that (a) is true of them.

The nice thing about hypotheses like H where every putative cause has an open back path is that we can tell by inspection whether H is true or not. For no x_j can appear in a functionally correct equation with a causally wrong coefficient unless some factor appears on the right-hand side of that equation along with a factor from its back path.[9] But according to (a), no factor from the back path of q_j can appear as a cause of q_e in the same law as q_j. The equation for x_e is thus a true causal law, so long as nothing appears on the right-hand side that is from the back path of any other factor that appears there. Given (b), we can tell this just by looking. According to Cartwright (l989), J. L. Mackie's famous example of the London workers and the Manchester hooters works in just this way.

3) Randomized treatment/control experiments are the gold standard for establishing causal laws in areas where we do not have sufficient knowl-

[8] Thanks to David Danks for highlighting this feature.

[9] The proof is similar to the proof of the theorems in Section 2 above. See Cartwright 1989. (Note that the argument in Spirtes, Glymour, and Scheines 1993 against this result uses as a putative counterexample one that does not meet the conditions set.)

https://doi.org/10.1086/367876 Published online by Cambridge University Press

edge to control confounding factors directly. These experiments require that there be some method for varying the causal factors under test without in any other way producing variation in the effect in question. In an epistemologically convenient system, the exogenous quantities peculiar to each factor provide just such a method.

3.2. The Second Definitions. Now for "intervening." The idea is to "vary" the value of the targeted quantity by adjusting its exogenous cause in just the right way, keeping fixed the values of all the other exogenous causes. But as I indicated, neither the idea of our manipulating nor of our varying anything matters. All we need is to consider what would happen were two different values of the exogenous cause of the targeted quantity to occur in two otherwise identical situations. So I propose the following definition: A *variation/intervention of values* is a calculation of $\Delta_j^\alpha q_k$ for some j, k, α. Direct inspection of the reduced form for q_k shows the following to hold:

Lemma (on reduced forms and causality): If q_j does not cause q_k then $\Delta_j^\alpha q_k = 0$.

Along with the notion of "intervention," we have to introduce new notions of invariance and causal correctness as well, otherwise the kinds of theorems we are interested in will not follow. The result in one direction still follows: any causally correct equation will be invariant under variation/intervention. But that is because *any* true equation will be, including all those equations that suggest joint effects of a common cause as causes of each other. Hence the result we really want in order to test for causal correctness will not follow, i.e., it is not true that any equation that is invariant under value variation/intervention will be causally correct (even if we restrict attention, as below, to equations in which no right-hand-side quantity causes any other).

What notion shall we substitute for that of invariance? The answer must clearly be tied to what kind of causal claim is made since we are not, after all, interested in invariance itself but pursue it as a test for causality. So far the kind of causal claim we have considered is terrifically restricted given our usual epistemic position. For we consider only hypotheses that claim to offer a complete (i.e., determining) set of causes and with exactly the weights nature assigns them. One way to be less demanding would be to ask for causes but not insist on weights.

Another alternative is to insist that the weights be correct, but not insist on a complete set of causes. This is the one I consider here. If we are offering claims with some causes omitted, what form should the hypotheses take? One standard answer is that they take the form of *regression equations:*

https://doi.org/10.1086/367876 Published online by Cambridge University Press

$$\textbf{R:}\ x_k\ c = \Sigma a_{kj}x_j + \Psi_k,\ \text{for}\ \Psi_k \perp x_j\ \text{for all j},$$

where $x \perp y$ means that $<xy> - <x><y> = 0$. This of course only makes sense if there is a probability measure from which the expectations are derived. So the use of hypotheses of this form involves an additional restriction on the kinds of systems under study, as follows. An *Epistemically Convenient Linear Deterministic System with Probability Measure (ECLDSwPM)* is an epistemically convenient linear deterministic system that satisfies A_8.

> A_8: *Existence of a probability measure.* The quantities in Q can be represented by random variables $x_1, \ldots, x_m$ which have a probability measure defined over them. (Following conventional notation, we can relabel the x's just as we have the q's so that $\{x_1, \ldots, x_m\} = \{x_1, \ldots, x_{m/2}, u_1, \ldots, u_{m/2}\}$).

What does an equation of form R assert? This kind of equation is often on offer but generally without any explanation about what claims it is supposed to make. I take it that it is supposed to include only genuine causes of x_k and moreover to tell us the correct weights of these. I propose, therefore, to define *correctness* thus: an equation of the form R: $x_k\ c = \Sigma a_{kj}x_j + \Psi_k\ (1 \leq j \leq m/2)$, for $\Psi_k \perp x_j$, is *correct* iff there exist $\{b_j\}$ (possibly $b_j = 0$), $\{q_j'\}$ such that $q_k\ c = \Sigma a_{kj}q_j + \Sigma b_j q_j' + u_k (1 \leq j \leq m/2)$, where q_j does not cause q_j'. This last restriction ensures that all omitted factors are causally antecedent to or "simultaneous" with those mentioned in the regression formula.

It may be useful to consider an example:

$$q_1\ c = u_1$$

$$q_2\ c = a_{21}q_1 + u_2$$

$$q_3\ c = u_3$$

$$q_4\ c = a_{41}q_1 + a_{42}q_2 + a_{43}q_3 + u_4$$

In this causal system the equation

$$x_4\ c = (a_{41} + a_{42}a_{21})q_1 + R$$

is correct. It may seem worrying that q_2 is omitted from the right-hand side of the regression equation and it is caused by q_1, which is included. But this is all right. The claims of the regression equation are correct under the proposed definition because there is a true causal law in which the

https://doi.org/10.1086/367876 Published online by Cambridge University Press

coefficient of q_1 is that given in the regression equation, and no factors in the true law that do not appear in the regression equation are caused by ones that are mentioned.

Now return to the unresolved issue of what can be introduced in place of invariance to dovetail with this characterization of *correctness* for regression equations. As I indicated in the Introduction, the notion that I use is not a notion of invariance at all. It is rather a notion of correct prediction: correct prediction of variation in values as situations vary in specific ways. This is not in any way a new notion, but it is one that Woodward has recently directed our attention to and that he has developed at length. I believe that what I define here is the right way to characterize his ideas when applied to epistemically convenient linear deterministic systems, and I take it that the theorem I prove is one precise formulation of what he argues for (once a number of caveats are added to his claims).

What do equations of form R predict about the difference in the size of effect between these two situations? If R's claims are correct, the difference in the effect given a variation of the special exogenous variable that causes one of the right-hand-side variables, say x_J, should be thus: $\Delta_J^\alpha q_k = \Sigma a_{kj}\Delta_J^\alpha q_j + \Sigma b_j\Delta_J^\alpha q_j'$ for some $\{b_j\}$ and $\{q_j'\}$, where no q_j causes any q_j'. By inspection of the reduced form equations in an ECLDSwPM, we see that the second term on the right-hand side is zero, since q_J does not cause any of the quantities that appear there. So R's predictions are correct just in case $\Delta_J^\alpha q_k = \Sigma a_{kj}\Delta_J^\alpha q_j$. So let us define: an equation of form R *correctly predicts first differences for all right-hand-side variables* if and only if, $\Delta_J^\alpha q_k = \Sigma a_{kj}\Delta_J^\alpha q_j$ for all α and for all J, where J ranges over the right-hand-side variables.

3.3. The Second Theorems. Now I can state the relevant theorem:

> *Theorem 2a.* A regression equation for q_k, x_k c $= \Sigma_{j=1}{}^{k-1} a_{kj}x_j + \Psi_k$, is causally correct iff for all α and for all J, $1 \leq J \leq k-1$, $\Delta_J^\alpha q_k = \Sigma a_{kj}\Delta_J^\alpha$ q_j; i.e., iff the equation predicts rightly the first differences in q_k generated from any value variation/intervention in any right-hand-side variable.

First a note on notation. In general there will be more q's in the underlying causal system than are represented by x's from the causal structure. For convenience I suppose that the q's are ordered following the x's: i.e., q_j is the quantity represented by x_j. This means that we cannot presuppose that q_i is causally prior to q_{i+1}.

Proof of Theorem 2a. The proof from correctness to the prediction of first differences in q_k under variations of right-hand-side variables is trival. To go the other direction, first reorder the *q*'s so that they are numbered in their true causal order (so, q_j can only cause q_{j+1} for $l \geq 1$), which we can do without commitment since the ordering is arbitrary to begin with. Then renumber the *x*'s accordingly. For all $1 \leq J \leq k-1$ and all α we suppose that

$$\Delta_J^\alpha q_k = \sum_{i=1}^{k-1} a_{ki}\Delta_J^\alpha q_i$$

Note first that we can always write

$$q_k = \sum_{i=1}^{k-1} A_{ki}q_i + \sum_{j=k+1}^{m/2} B_{kj}q_j + u_k$$

where q_j, $k+1 \leq j \leq m/2$, is not caused by q_i, $1 \leq i \leq k-1$, with A_{ki} possibly 0. For consider any causal equation of this form where some of the q_j are caused by some of the q_i. To find a true causal law of the required form simply substitute for each of the unwanted q_j an expansion in a set of causes of q_j, all of which occur prior to all q_i. From this it follows from our lemma that for all J such that $i \leq J \leq k-1$,

$$\Delta_J^\alpha q_k = \sum_{i=1}^{k\text{-}1} A_{ki}\Delta_J^\alpha q_i.$$

We need to show that $A_{ki} = a_{ki}$. Consider first $\Delta_L^\alpha q_k$, where $1 \leq L \leq k-1$ and q_L is causally posterior to all other q_i for $1 \leq i \leq k-1$:

$$\Delta_L^\alpha q_k = a_L\Delta_L^\alpha q_L = A_{kL}\Delta_L^\alpha q_L,$$

where the first equality comes from the assumption that the equation for q_k predicts first differences correctly and the second from the true law for q_k. It follows that $a_{kL} = A_{kL}$.

Next consider $\Delta_{L'}^\alpha q_k$, where $i \leq L' \leq k-1$ and $q_{L'}$ is causally posterior to all other q_i for $1 \leq i \leq k-1$ except for L.

$$\Delta_{L'}^\alpha q_k = a_{kL}\Delta_{L'}^\alpha q_L + a_{kL'}\Delta_{L'}^\alpha q_{L'} = A_{kL}\Delta_{L'}^\alpha q_L + A_{kL'}\Delta_{L'}^\alpha q_{L'}$$

for the same reasons as before. Since $a_{kL} = A_{kL}$, it follows that $a_{kL'} = A_{kL'}$. And so on for each coefficient in turn. Q.E.D.

Notice, however, that this theorem is not very helpful because it will be hard to tell whether an equation has indeed predicted first differences rightly. That is because we will not know what $\Delta_J^\alpha q_j$ should be unless we know how variations in u_J affect q_j and to know that we will have to know

https://doi.org/10.1086/367876 Published online by Cambridge University Press

the causal relations between q_J and q_j. So in order to judge whether each of the q_j affects q_k in the way hypothesized, we will have to know already how they affect each other. If we happen to know that none of them affect the others at all, we will be in a better situation, since the following can be trivially derived from Theorem 2a:

> *Theorem 2b.* A regression equation $x_k\ c = \Sigma_{j=1}{}^{k-1}a_{kj}x_j + \Psi_k$ in which no right-hand side variable causes any other is causally correct iff for all α and J, $\Delta^{\alpha}_{J}q_k = a_{kJ}\Delta^{\alpha}_{J}u_J$.

We can also do somewhat better if we have a complete set of hypotheses about the right-hand-side variables. To explain this, let me define a *complete causal structure* that represents an ECLDSwPM , $<Q = \{q_1, \ldots, q_{m/2}, u_1, \ldots, u_{m/2}\}$, CL> as a pair $<X = \{x_1, \ldots, x_n: 1 \leq n \leq m/2\}$, μ, CLH>, where μ is a probability measure over the x's and where the causal law hypotheses, CLH, have the following form:

$$x_1\ c = \Psi_1$$

$$x_2\ c = a_{21}x_1 + \Psi_2$$

$$\vdots$$

$$x_n\ c = \Sigma^{n-1}_{j=1}a_{nj}x_j + \Psi_n,$$

where $\Psi_j \perp x_k$, for all $k < j$. In general $n < m/2$. Since the ordering of the q's has no significance, we will again suppose that they are ordered so that q_j is represented by x_j. Now I can formulate

> *Theorem 2c.* If for all x_k in a complete causal structure, $\Delta^{\alpha}_{J}q_k = \Delta^{\alpha}_{J}x_k$ *as predicted by the causal structure* for all α and J, $1 \leq J \leq n$, then all the hypotheses of the structure are correct.

For the proof we need some notation and a convention. What does the causal structure predict about differences in q_k for $\Delta^{\alpha}_{k}u_k$? I take it to predict that $\Delta^{\alpha}_{k}q_k = \Delta^{\alpha}_{k}u_k = \alpha$. To denote a *predicted* difference I use Δ', with Δ reserved for real differences (i.e., those that follow from the causal system being modeled in the causal structure). So the antecedent of Theorem 2c thus requires that for all J, $1 \leq J \leq n$, $\Delta^{\alpha}_{J}q_k = \Delta'^{\alpha}_{J'}x_k$.

Proof of Theorem 2c. Consider the *k*th equation in the structure

$$x_k\ c = \sum_{i=1}^{k-1} a_{ki}x_i + \Psi_k$$

We need to show that

https://doi.org/10.1086/367876 Published online by Cambridge University Press

$$q_k \; c = \sum_{i=1}^{k-1} a_{ki} q_i + \sum_{j=k+1}^{m/2} b_{kj} q_j + u_k$$

where q_i does not cause q_j for $1 \leq i \leq k-1$ and $k+1 \leq j \leq m/2$. We know that for some $\{A_{ki}\}$, $\{B_{ki}\}$

$$q_k \; c = \sum_{i=1}^{k-1} A_{ki} q_i + \sum_{j=k+1}^{m/2} B_{kj} q_j + u_k$$

where q_i does not cause q_j for $1 \leq i \leq k-1$ and for j such that $k+1 \leq j \leq m/2$ and $B_{kj} \neq 0$. So we need to establish that there is a set of A_{ki} such that $A_{ki} = a_{ki}$ for all i such that $1 \leq i \leq k-1$. We do so by backwards induction: show first that the coefficient of x_{k-1} is correct and work backwards from there. Note for the proof that since q_i, $1 \leq i \leq k-1$, does not cause q_j, for any j such that $k-1 \leq j \leq m/2$ and $B_{kj} \neq 0$, $\Delta_i^{\alpha} \sum_{j=k+1}^{m/2} B_{kj} q_j = 0$ for $1 \leq i \leq k-1$.

Inductive Base. To show,

$$\Delta_{k-1}^{\alpha} q_k = \sum_{i=1}^{k-1} A_{ki} \Delta_{k-1}^{\alpha} q_i = \sum_{i=1}^{k-1} A_{ki} \Delta_{k-1}'^{\alpha} q_i = A_{kk\text{-}1} \alpha$$
$$= \Delta_{k-1}'^{\alpha} q_k = \sum_{i=1}^{k-1} a_{ki} \Delta_{k-1}'^{\alpha} q_i = A_{ki} \alpha$$

So $A_{kk-1} = a_{kk-1}$.

Inductive Argument. Given $A_{k,p+s} = a_{k,p+s}$ for $1 \leq s < k-1-p$, to show $A_{kp} = a_{kp}$, consider what happens given Δ_p^{α}. Using the reduced form for q_i plus the assumption that all first difference predictions are right, and the fact that $\Delta'^{\alpha}_{p} q_i = 0$ for $i<p$, we have

$$\Delta_p^{\alpha} q_k = \sum_{i=1}^{k-1} A_{ki} \Delta_p^{\alpha} q_i = \sum_{i=1}^{k-1} A_{ki} \Delta_p'^{\alpha} q_i = \sum_{i=p}^{k-1} A_{ki} \Delta_p'^{\alpha} q_i$$
$$= \sum_{i=p}^{k-1} A_{ki} \Delta_p'^{\alpha} u_p \sum_{l=p}^{i-1} a_{il} \sum_{m=p}^{l-1} a_{lm} \ldots$$
$$= A_{kp} \alpha + \sum_{i=p+1}^{k-1} A_{ki} \alpha \sum_{l=p}^{i-1} a_{il} \sum_{m=p}^{l-1} a_{lm} \ldots$$
$$= \Delta_p'^{\alpha} q_k = \sum_{i=p}^{k-1} a_{ki} \Delta_p'^{\alpha} q_i = a_{kp} \alpha + \sum_{i=p+1}^{k-1} a_{ki} \alpha \sum_{l=p}^{i-1} a_{il} \sum_{m=p}^{l-1} a_{lm} \ldots$$

https://doi.org/10.1086/367876 Published online by Cambridge University Press

By hypotheses of the induction $A_{ki} = a_{ki}$, for $p+1 \leq i \leq k-1$. Hence $A_{kp} = a_{kp}$.

There is one important point about exogenous variables that we need to be clear about to understand the significance of the theorems. By definition, $\Delta_J^{\alpha} q$ is the difference in q given a difference in u_J with all other exogenous quantities *in the system,* not just those *in the structure,* held fixed. It is easy to see why. Consider a six-quantity system

$$q_3 \ c = u_3$$

$$q_1 \ c = a_{13} q_3 + u_1$$

$$q_2 \ c = a_{23} q_3 + u_2$$

and a two-variable causal structure to represent it

$$x_1 \ c = \Psi_1$$

$$x_2 \ c = c_{21} \ x_1 + \Psi_2 .$$

These will be true viewed just as regression equations given

$$c_{21} = a_{23} a_{13}/[1 + a_{13}^2] \text{ and } \Psi_2 = \left(a_{23}/a_{13} - a_{23} a_{13}/[1 + a_{13}^2]\right) q_1 + u_2 - (a_{23}/a_{13}) u_1 .^{10}$$

If u_1 varies while u_2 and u_3 do not, then we will see rightly that the equation for x_2 is not correct. But if as u_1 varies, u_3 varies as well in such a way that $a_{23} \Delta u_3 = c_{21} \Delta u_1$, then the equation for x_2 will produce the right first difference predictions for x_2. That is why, to get a proper test for the equation, we must consider variation in exogenous variables in the structure while all other exogenous quantities in the system and (also in the structure) remain constant.

This makes the results more difficult to put to use than we might have hoped. In the first place, for the theorems to apply at all, we need to know that we are dealing with an epistemically convenient system—one for which the exogenous factors have no cross restraints. But it is hard enough to know about the cross restraints on the exogenous causes for a set of putative causes we are considering in our structure, let alone for a lot of possible causes in the system that we have no idea of.

[10] Recall that for $x_2 = c_{21} x_1 + \Psi_2$ to be a regression equation, $\langle x_1, \Psi_2 \rangle = 0$. I assume here that the u's have mean 0, variance 1 and $\langle u_i, u_j \rangle = 0$, $i \neq j$.

https://doi.org/10.1086/367876 Published online by Cambridge University Press

Suppose, though, that we do have good reason to think that the system we are studying is epistemically convenient (or we are prepared to bet on it). How would we use the theorems to which that entitles us? The most straightforward application of the theorems to test a hypothesis about the causes of q would consider variations in the exogenous factors for q's putative causes holding fixed all other exogenous factors, where these have to include all other exogenous factors *in the system.* So we would have to know what these factors are. Again, it is hard enough to know what the exogenous causes are for factors we can identify without having to know what they are for factors we do not know about.[11]

I take it that this is the chief motivation for stressing manipulation. It seems that if we vary the putative causes *at will* or *arbitrarily* the variation will not match any natural variation in other exogenous factors. But we know that is not true. Coincidences happen, even when the variation is chosen completely arbitrarily—which we know at any rate is hard to achieve due to placebo effects, experimenter bias, and the like. For these theorems, exactly what is required is the right kind of variation, no more and no less. So the emphasis on manipulation for invariance tests of causality is misplaced, except as a not-100%-reliable methodological tool.[12]

4. Final Remark. We are interested in whether invariance (or some substitute) under intervention is a sure sign of correctness in a causal claim. I have formalized two distinct senses commonly in use for each of the three concepts involved. That means there are eight versions of the question using just the concepts defined here. I have answered the question for only three: (1) for invariance under causal-law variation and correctness *simpliciter,* the answer (with caveats) is *yes;* (2) for invariance under intervention/variation of values and correctness *simpliciter,* the answer is *no;* and (3) for prediction of first differences under intervention/variation of values, the answer for prediction of first differences is *yes.*

Clearly we can carry on pursuing the other combinations, or devise modifications of the concepts that might serve better in hunting good tests. With respect to the concepts deployed here, one in particular is fairly central: that is the version of the question involving parameter correctness under first difference prediction. That's because of our usual epistemic situation. First, when a hypothesis does not involve a full set of determin-

[11] As we know, randomized treatment/control experiments are designed to allow us to get around our lack of knowledge of the exogenous factors for missing factors. But the knowledge that we have succeeded in the aims of randomizing even when we have used our best methods is again hard to come by.

[12] As, of course, is widely recognized in the experimental literature in the social sciences.

https://doi.org/10.1086/367876 Published online by Cambridge University Press

ing factors, we are forced to look at the predictions about first differences since it makes no sense to ask whether the hypothesis is invariant or not; and correlatively, we can demand only correctness in the parameters on offer, not full correctness. Second, when the system under study is not epistemologically convenient, we are forced to use causal-law intervention to get the variation we need. I take it the answer for this particular combination is *yes*—with caveats. But, as with any answer, we need a clear statement of the caveats and a convincing proof.

There is a division among philosophers of science between those who believe that formalization is essential to understanding and those who do not. Here I have been arguing on the side of the formalizers. For me the point of studying the relations between causality and invariance is to make better causal judgments; and if different ways of making our theses precise matter to how we make our judgements, then we had better be precise. We have seen that they do matter. Invariance under intervention is a fine test for causality if the intervention involves looking at what happens in different causal systems, but not if it involves looking at different situations governed by the same system of laws. Or, when we do look at different situations, what counts as a test of a causal hypothesis when none of the putative causes cause any of the others will not serve when some do cause others.

Formalization is, however, nowhere near the end of the road. We still face the traditional problem of what all these precisely defined concepts mean in full empirical reality. In particular what is the difference between a variation in the value of a putative cause that arises from a variation in the causal system governing it versus one that arises from a variation in an exogenous cause that operates within the original system? Imagine I am about to do a randomized treatment-control experiment. How do I judge whether my proposed method of inducing the treatment fits one description or the other? I do not know how to answer the question. Perhaps indeed the distinction, which makes such clear sense conceptually, does not fit onto the empirical world it is intended to help us with. Formalization is, to my mind, the easy (though necessary) part of the job. Our next task is to provide an account of the connection between our formal concepts and what we can do in practice.

REFERENCES

Balke, A., and Judea Pearl (1995), "Counterfactuals and Policy Analysis in Structural Models", in P. Besnard and S. Hanks (eds.), *Uncertainty in Artificial Intelligence* 11, San Francisco, CA: Morgan Kaufmann, 11–18.

Cartwright, Nancy (1989), *Nature's Capacities and Their Measurement.* Oxford: Oxford University Press.

——— (1999), *The Dappled World.* Cambridge: Cambridge University Press.

——— (2000), *Measuring Causes: Invariance, Modularity and the Causal Markov Condition.* Measurement in Physics and Economics Discussion Paper Series, LSE, London.

https://doi.org/10.1086/367876 Published online by Cambridge University Press

Engle, Robert, David Hendry, and Jean Richard (1983), "Exogeneity", *Econometrica* 51: 277–304.

Glymour, Clark, Richard Scheines, Peter Spirtes, and Kevin Kelly (1987), *Discovering Causal Structure.* New York: Academic Press.

Hausman, Daniel (1998), *Causal Asymmetries.* Cambridge: Cambridge University Press.

——— and James Woodward (1999), "Independence, Invariance, and the Causal Markov Condition", *British Journal for the Philosophy of Science* 50: 521–583.

Hoover, Kevin (2001), *Causality in Macroeconomics.* Cambridge: Cambridge University Press.

Pearl, Judea (2000a), "The Logic of Counterfactuals in Causal Inference (Discussion of "Causal Inference without Counterfactuals")", *Journal of the American Statistical Association* Spring.

——— (2000b), *Causality.* Cambridge: Cambridge University Press.

Price, Huw (1991), "Agency and Probabilistic Causality", *British Journal for the Philosophy of Science* 42: 157–176.

Redhead, Michael (1987), *Incompleteness, Nonlocality, and Realism: A Prolegomenon to the Philosophy of Quantum Mechanics,* Oxford: Oxford University Press.

Spirtes, Peter, Clark Glymour, and Richard Scheines (1993), *Causation, Prediction, and Search.* New York: Springer Verlag.

Woodward, James (1997), "Explanation, Invariance, and Intervention", *Philosophy of Science* 64 (Proceedings): S26-S41.

https://doi.org/10.1086/367876 Published online by Cambridge University Press