CHAPTER 10

CAUSATION AND EXPLANATION IN SOCIAL SCIENCE

HENRY E. BRADY

1 Causality

Humans depend upon causation all the time to explain what has happened to them, to make realistic predictions about what will happen, and to affect what happens in the future. Not surprisingly, we are inveterate searchers after causes. Almost no one goes through a day without uttering sentences of the form *X caused Y* or *Y occurred because of X*. Causal statements explain events, allow predictions about the future, and make it possible to take actions to affect the future. Knowing more about causality can be useful to social science researchers.

Philosophers and statisticians know something about causality, but entering into the philosophical and statistical thickets is a daunting enterprise for social scientists because it requires technical skills (e.g. knowledge of modal logic) and technical information (e.g. knowledge of probability theory) that is not easily mastered. The net payoff from forays into philosophy or statistics sometimes seems small compared to the investment required. The goal of this chapter is to provide a user-friendly synopsis of philosophical and statistical musings about causation. Some technical issues will be discussed, but the goal will always be to ask about the bottom line—how can this information make us better researchers?

Three types of intellectual questions typically arise in philosophical discussions of causality:

- *Psychological and linguistic*—What do we *mean* by causality when we use the concept?
- *Metaphysical or ontological*—What *is* causality?
- *Epistemological*—How do we *discover* when causality is operative?[1]

Four distinct approaches to causality, summarized in Table 10.1, provide answers to these and other questions about causality.[2] Philosophers debate which approach is the right one. For our purposes, we embrace them all. Our primary goal is developing better social science methods, and our perspective is that all these approaches capture some aspect of causality. Therefore, practical researchers can profit from drawing lessons from each one of them even though their proponents sometimes treat them as competing or even contradictory. Our standard has been whether or not we could think of concrete examples of research that utilized (or could have utilized) a perspective to some advantage. If we could think of such examples, then we think it is worth drawing lessons from that approach.

A really good causal inference should satisfy the requirements of all four approaches. Causal inferences will be stronger to the extent that they are based upon finding all the following: (1) Constant conjunction of causes and effects required by the neo-Humean approach. (2) No effect when the cause is absent in the most similar world to where the cause is present as required by the counterfactual approach. (3) An effect after a cause is manipulated. (4) Activities and processes linking causes and effects required by the mechanism approach.

The claim that smoking causes lung cancer, for example, first arose in epidemiological studies that found a correlation between smoking and lung cancer. These results were highly suggestive to many, but this correlational evidence was insufficient to others (including one of the founders of modern statistics, R. A. Fisher). These studies were followed by experiments that showed that, at least in animals, the absence of smoking reduced the incidence of cancer compared to the incidence with smoking when similar groups were compared. But animals, some suggested, are not people. Other studies showed that when people stopped smoking (that is, when the putative cause of cancer was manipulated) the incidence of cancer went down as well. Finally, recent studies have uncovered biological mechanisms that explain the link between smoking and lung cancer. Taken together the evidence for a relationship between smoking and lung cancer now seems overwhelming.

[1] A fourth question is pragmatic: How do we *convince* others to accept our explanation or causal argument? A leading proponent of this approach is Bas van Fraassen (1980). Kitcher and Salmon (1987, 315) argue that "van Fraassen has offered the best theory of the pragmatics of explanation to date, but . . . if his proposal is seen as a pragmatic theory of explanation then it faces serious difficulties" because there is a difference between "a theory of the pragmatics of explanation and a pragmatic theory of explanation." From their perspective, knowing how people convince others of a theory does not solve the ontological or epistemological problems.

[2] Two important books on causality are not covered in this chapter, although the author has profited from their insights. Pearl (2000) provides a comprehensive approach to causality rooted in a Bayesian perspective. Shafer (1996) links decision theory and causal trees in a novel and useful way.

Table 10.1. Four approaches to causality

	Neo-Humean regularity	Counterfactual	Manipulation	Mechanisms and capacities
Major authors associated with the approach	Hume (1739); Mill (1888); Hempel (1965); Beauchamp and Rosenberg (1981)	Weber (1906); Lewis (1973b; 1973c; 1986)	Gasking (1955); Menzies and Price (1993); von Wright (1971)	Harre and Madden (1975); Cartwright (1989); Glennan (1996);
Approach to the symmetric aspect of causality	Observation of constant conjunction and correlation	Truth in otherwise similar worlds of "if the cause occurs then so does the effect" and "if the cause does not occur then the effect does not occur"	Recipe that regularly produces the effect from the cause	Consideration of whether there is a mechanism or capacity that leads from the cause to the effect
Approach to the asymmetric aspect of causality	Temporal precedence	Consideration of the truth of: "if the effect does not occur, then the cause may still occur"	Observation of the effect of the manipulation	An appeal to the operation of the mechanism
Major problems solved	Necessary connection	Singular causation; nature of necessity	Common cause and causal direction	Pre-emption
Emphasis on causes of effects or effects of causes?	Causes of effects (e.g. focus on dependent variable in regressions.)	Effects of causes (e.g. focus on treatment's effects in experiments)	Effects of causes (e.g. focus on treatment's effects in experiments)	Causes of effects (e.g. focus on mechanism that creates effects)
Studies with comparative advantage using this definition	Observational and causal modelling	Experiments; case study comparisons; counterfactual thought experiments	Experiments; natural experiments; quasi-experiments	Analytic models; case studies

2 Counterfactuals

Causal statements are so useful that most people cannot let an event go by without asking why it happened and offering their own "because." They often enliven these discussions with counterfactual assertions such as "if the cause had not occurred, then the effect would not have happened." A counterfactual is a statement, typically in the subjunctive mood, in which a false or "counter to fact" premise is followed by some assertion about what would have happened if the premise were true. For example, the butterfly ballot was used in Palm Beach County Florida in 2000 and George W. Bush was elected president. A counterfactual assertion might be "if the butterfly ballot had not been used in Palm Beach County in 2000, then George Bush would not have been elected president." The statement uses the subjunctive ("if the butterfly ballot had not been used,...then George Bush would not have been elected"), and the premise is counter to the facts. The premise is false because the butterfly ballot was used in Palm Beach County in the real world as it unfolded. The counterfactual claim is that without this ballot, the world would have proceeded differently, and George Bush would not have been president. Is this true?

The truth of counterfactuals is closely related to the existence of causal relationships. The counterfactual claim made above implies that there is a causal link between the butterfly ballot (the cause X) and the election of George Bush (the effect Y). The counterfactual, for example, would be true if the butterfly ballot *caused* Al Gore to lose enough votes so that Bush was elected. Then, if the butterfly ballot had not been used, Al Gore would have gotten more votes and won the election.

Another way to think about this is to simply ask what would have happened in the *most similar world* in which the butterfly ballot was not used. Would George Bush still be president? One way to do this would be to rerun the world with the cause eradicated so that the butterfly ballot was not used. The world would otherwise be the same. If George Bush did not become president, then we would say that the counterfactual is true. Thus, the statement that the butterfly ballot *caused* the election of George W. Bush is essentially the same as saying that in the *most similar world* in which the butterfly ballot did not exist, George Bush would have lost. The existence of a causal connection can be checked by determining whether or not the counterfactual would be true in the most similar possible world where its premise is true. The problem, of course, is defining the most similar world and finding evidence for what would happen in it.

Beyond these definitional questions about most similar worlds, there is the problem of finding evidence for what would happen in the most similar world. We cannot rerun the world so that the butterfly ballot is not used. What can we do? Many philosophers have wrestled with this question, and we discuss the problem in detail later in the section on the counterfactual approach to causation.[3] For now, we merely

[3] Standard theories of logic cannot handle counterfactuals because propositions with false premises are automatically considered true which would mean that all counterfactual statements, with their false

note that people act as if they can solve this problem because they assert the truth of counterfactual statements all the time.

3 Exploring Three Basic Questions about Causality

Causality is at the center of explanation and understanding, but what, exactly, is it? And how is it related to counterfactual thinking? Somewhat confusingly, philosophers mingle psychological, ontological, and epistemological arguments when they discuss causality. Those not alerted to the different purposes of these arguments may find philosophical discussions perplexing as they move from one kind of discussion to another. Our primary focus is epistemological. We want to know when causality is truly operative, not just when some psychological process leads people to believe that it is operative. And we do not care much about metaphysical questions regarding what causality really is, although such ontological considerations become interesting to the extent that they might help us discover causal relationships.

3.1 Psychological and Linguistic Analysis

Although our primary focus is epistemological, our everyday understanding, and even our philosophical understanding, of causality is rooted in the psychology of causal inference. Perhaps the most famous psychological analysis is David Hume's investigation of what people mean when they refer to causes and effects. Hume (1711–76) was writing at a time when the pre-eminent theory of causality was the existence of a necessary connection—a kind of "hook" or "force"—between causes and their effects so that a particular cause must be followed by a specific effect. Hume looked for the feature of causes that guaranteed their effects. He argued that there was no evidence for the necessity of causes because all we could ever find in events was the contiguity, precedence, and regularity of cause and effect. There was no evidence for any kind of hook or force. He described his investigations as follows in his *Treatise of Human Nature* (1739):

> *What is our idea of necessity, when we say that two objects are necessarily connected together?* . . . I consider in what objects necessity is commonly supposed to lie; and finding that it is always ascribed to causes and effects, I turn my eye to two objects supposed to be placed in that

premises, would be true, regardless of whether or not a causal link existed. Modal logics, which try to capture the nature of necessity, possibility, contingency, and impossibility, have been developed for counterfactuals (Lewis 1973*a*; 1973*b*). These logics typically judge the truthfulness of the counterfactual on whether or not the statement would be true in the most similar possible world where the premise is true. Problems arise, however, in defining the most similar world.

relation, and examine them in all the situations of which they are susceptible. I immediately perceive that they are *contiguous* in time and place, and that the object we call cause *precedes* the other we call effect. In no one instance can I go any further, nor is it possible for me to discover any third relation betwixt these objects. I therefore enlarge my view to comprehend several instances, where I find like objects always existing in like relations of contiguity and succession. The reflection on several instances only repeats the same objects; and therefore can never give rise to a new idea. But upon further inquiry, I find that the repetition is not in every particular the same, but produces a new impression, and by that means the idea which I at present examine. For, after a frequent repetition, I find that upon the appearance of one of the objects the mind is *determined* by custom to consider its usual attendant, and to consider it in a stronger light upon account of its relation to the first object. It is this impression, then, or *determination*, which affords me the idea of necessity. (Hume, 1978 [1739], 155)[4]

Thus for Hume the *idea* of necessary connection is a psychological trick played by the mind that observes repetitions of causes followed by effects and then presumes some connection that goes beyond that regularity. For Hume, the major feature of causation, beyond temporal precedence and contiguity, is simply the regularity of the association of causes with their effects, but there is no evidence for any kind of hook or necessary connection between causes and effects.[5]

The Humean analysis of causation became the predominant perspective in the nineteenth and most of the twentieth century, and it led in two directions, both of which focused upon the logical form of causal statements. Some, such as the physicist Ernst Mach, the philosopher Bertrand Russell, and the statistician/geneticist Karl Pearson, concluded that there was nothing more to causation than regularity so that the entire concept should be abandoned in favor of functional laws or measures of association such as correlation which summarized the regularity.[6] Others, such as the philosophers John Stuart Mill (1888), Karl Hempel (1965), and Tom Beauchamp and

[4] In the *Enquiry* (1748, 144–5) which is a later reworking of the *Treatise*, Hume says: "So that, upon the whole, there appears not, throughout all nature, any one instance of connexion, which is conceivable by us. All events seem entirely loose and separate. One event follows another; but we never can observe any tye between them. They seem *conjoined*, but never *connected*. And as we can have no idea of any thing, which never appeared to our outward sense or inward sentiment, the necessary conclusion *seems* to be, that we have no idea of connexion or power at all, and that these words are absolutely without meaning, when employed either in philosophical reasonings, or common life. . . . This connexion, therefore, we feel in the mind, this customary transition of the imagination from one object to its usual attendant, is the sentiment or impression, from which we form the idea of power or necessary connexion."

[5] There are different interpretations of what Hume meant. For a thorough discussion see Beauchamp and Rosenberg (1981).

[6] Bertrand Russell famously wrote that "the word 'cause' is so inextricably bound up with misleading associations as to make its complete extrusion from the philosophical vocabulary desirable. . . . The law of causality, like so much that passes muster among philosophers, is a relic of a bygone age, surviving like the monarchy, only because it is erroneously supposed to do no harm" (Russell 1918). Karl Pearson rejected causation and replaced it with correlation: "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of even modern science, namely the category of cause and effect. Is this category anything but a conceptual limit to experience, and without any basis in perception beyond a statistical approximation?" (Pearson 1911, vi). "It is this conception of correlation between two occurrences embracing all relationship from absolute independence to complete dependence, which is the wider category by which we have to replace the old idea of causation" (Pearson 1911, 157).

Alexander Rosenberg (1981), looked for ways to strengthen the regularity condition so as to go beyond mere accidental regularities. For them, true cause and effect regularities must be unconditional and follow from some lawlike statement. Their neo-Humean approach improved upon Hume's approach, but as we shall see, there appears to be no way to define lawlike statements in a way that captures all that we mean by causality.

What, then, do we typically mean by causality? In their analysis of the fundamental metaphors used to mark the operation of causality, the linguist George Lakoff and the philosopher Mark Johnson (1980*a*; 1980*b*; 1999) describe prototypical causation as "the manipulation of objects by force, the volitional use of bodily force to change something physically by direct contact in one's immediate environment" (1999, 177). Causes bring, throw, hurl, propel, lead, drag, pull, push, drive, tear, thrust, or fling the world into new circumstances. These verbs suggest that causation is forced movement, and for Lakoff and Johnson the "Causation Is Forced Movement metaphor is in a crucial way constitutive of the concept of causation" (187). Causation as forceful manipulation differs significantly from causation as the regularity of cause and effect because forceful manipulation emphasizes intervention, agency, and the possibility that the failure to engage in the manipulation will prevent the effect from happening. For Lakoff and Johnson, causes are forces and capacities that entail their effects in ways that go beyond mere regularity and that are reminiscent of the causal "hooks" rejected by Hume, although instead of hooks they emphasize manipulation, mechanisms, forces, and capacities.[7]

"Causation as regularity" and "causation as manipulation" are quite different notions, but each carries with it some essential features of causality. And each is the basis for a different philosophical or everyday understanding of causality. From a psychological perspective, their differences emerge clearly in research done in the last fifteen years on the relationship between causal and counterfactual thinking (Spellman and Mandel 1999). Research on this topic demonstrates that people focus on different factors when they think causally than when they think counterfactually. In experiments, people have been asked to consider causal attributions and counterfactual possibilities in car accidents in which they imagine that they chose a new route to drive home and were hit by a drunk driver. People's *causal attributions* for these accidents tend to "focus on antecedents that general knowledge suggest would covary with, and therefore predict, the outcome (e.g., the drunk driver)," but *counterfactual thinking* focuses on controllable antecedents such as the choice of route (Spellman and Mandel 1999, 123). Roughly speaking, causal attributions are based upon a regularity approach to causation while counterfactual thinking is based upon a manipulation approach to causation. The regularity approach suggests that drunken drivers typically cause accidents but the counterfactual approach suggests that in this instance the person's

[7] As we shall show, two different approaches to causation are conflated here. One approach emphasizes agency and manipulation. The other approach emphasizes mechanisms and capacities. The major difference is the locus of the underlying force that defines causal relationships. Agency and manipulation approaches emphasize human intervention. Mechanism and capacity approaches emphasize processes within nature itself.

choice of a new route was the cause of the accident because it was manipulable by the person. The logic of causal and the logic of counterfactual thinking are so closely related that these psychological differences in attributions lead to the suspicion that both the regularity and the manipulation approaches tell us something important about causation.

3.2 Ontological Questions

Knowing how most people think and talk about causality is useful, but we are ultimately more interested in knowing what causality actually is and how we would discover it in the world. These are respectively ontological and epistemological questions.[8] Ontological questions ask about the characteristics of the abstract entities that exist in the world. The study of causality raises a number of fundamental ontological questions regarding the *things that are causally related* and the *nature of the causal relation*.[9]

What are the things, the "causes" and the "effects" that are linked by causation? Whatever they are, they must be the same things because causes can also be effects and vice versa. But what are they? Are they facts, properties, events, or something else?[10] The practicing researcher cannot ignore questions about the definition of events. One of the things that researchers must consider is the proper definition of an event,[11] and a great deal of the effort in doing empirical work is defining events suitably. Not surprisingly, tremendous effort has gone into defining wars, revolutions, firms, organizations, democracies, religions, participatory acts, political campaigns, and many other kinds of events and structures that matter for social science research. Much could be said about defining events, but we shall only emphasize that defining events in a useful fashion is one of the major tasks of good social science research.

A second basic set of ontological questions concern the nature of the causal relationship. Is causality different when it deals with physical phenomena (e.g. billiard

[8] Roughly speaking, philosophy is concerned with three kinds of questions regarding "what is" (ontology), "how it can be known" (epistemology), and "what value it has" (ethics and aesthetics). In answering these questions, twentieth-century philosophy has also paid a great deal of attention to logical, linguistic, and even psychological analysis.

[9] Symbolically, we can think of the causal relation as a statement XcY where X is a cause, Y is an effect, and c is a causal relation. X and Y are the things that are causally related and c is the causal relation. As we shall see later, this relationship is usually considered to be incomplete (not all X and Y are causally related), asymmetric for those events that are causally related (either XcY or YcX but not both), and irreflexive (XcX is not possible).

[10] Events are located in space and time (e.g. "the WWI peace settlement at Versailles") but facts are not ("The fact that the WW I peace settlement was at Versailles"). For discussions of causality and events see Bennett (1988) and for causality and facts see Mellors (1995). Many philosophers prefer to speak of "tropes" which are particularized properties (Ehring 1997). Some philosophers reject the idea that the world can be described in terms of distinct events or tropes and argue for events as enduring things (Harre and Madden 1975, ch. 6).

[11] A potpourri of citations that deal with the definition of events and social processes are Abbott (1983; 1992; 1995), Pierson (2004), Riker (1957), Tilly (1984)

balls hitting one another or planets going around stars) than when it deals with social phenomena (democratization, business cycles, cultural change, elections) that are socially constructed?[12] What role do human agency and mental events play in causation?[13] What can we say about the time structure and nature of causal processes?[14] Our attitude is that social science is about the formation of concepts and the identification of causal mechanisms. We believe that social phenomena such as the Protestant ethic, the system of nation states, and culture exist and have causal implications. We also believe that reasons, perceptions, beliefs, and attitudes affect human behavior. Furthermore, we believe that these things can be observed and measured.

Another basic question about the causal relation is whether it is deterministic or probabilistic. The classic model of causation is the deterministic, clockwork Newtonian universe in which the same initial conditions inevitably produce the same outcome. But modern science has produced many examples where causal relationships appear to be probabilistic. The most famous is quantum mechanics where the position and momentum of particles is represented by probability distributions, but many other sciences rely upon probabilistic relationships. Geneticists, for example, do not expect that couples in which all the men have the same height and all the women have the same height will have children of the same height. In this case, the same set of (observed) causal factors produce a probability distribution over possible heights. We now know that even detailed knowledge of the couple's DNA would not lead to exact predictions. Probabilistic causation, therefore, seems possible in the physical sciences, common in the biological sciences, and pervasive in the social sciences. Nevertheless, following the custom of a great deal of philosophical work, we shall start with a discussion of deterministic causation in order not to complicate the analysis.

3.3 Epistemological Questions

Epistemology is concerned with how we can obtain intellectually certain knowledge (what the Greeks called "episteme"). How do we figure out that *X* really caused *Y*? At the dinner table, our admonition not to reach across the table might be met with "I didn't break the glass, the table shook," suggesting that our causal explanation for the broken glass was wrong. How do we proceed in this situation? We would probably try to rule out alternatives by investigating whether someone shook the table, whether there was an earthquake, or something else happened to disturb the glass. The problem here is that there are many possibilities that must be ruled out, and what must be ruled out depends, to some extent, on our definition of causality.

[12] For representative discussions see Durkheim (1982), Berger and Luckman (1966), von Wright (1971), Searle (1997), Wendt (1999).

[13] See Dilthey (1961), von Wright (1971, ch. 1), Davidson (2001), Searle (1969), Wendt (1999).

[14] In a vivid set of metaphors, Pierson (2004) compares different kinds of social science processes with tornadoes, earthquakes, large meteorites, and global warming in terms of the time horizon of the cause and the time horizon of the impact. He shows that the causal processes in each situation are quite different.

Learning about causality, then, requires that we know what it is and that we know how to recognize it when we see it. The simple Humean approach appears to solve both problems at once. Two events are causally related when they are contiguous, one precedes another, and they occur regularly in constant conjunction with one another. Once we have checked these conditions, we know that we have a causal connection. But upon examination, these conditions are not enough for causality because we would not say that night causes day, even though day and night are contiguous, night precedes day, and day and night are regularly associated. Furthermore, simple regularities like this do not make it easy to distinguish cause from effect—after all, day precedes night as well as night preceding day so that we could just as well, and just as mistakenly, say that day causes night. Something more is needed.[15] It is this something more that causes most of the problems for understanding causation. John Stuart Mill suggested that there had to be an "unconditional" relationship between cause and effect and modern neo-Humeans have required a "lawlike" relationship, but even if we know what this means[16] (which would solve the ontological problem of causation) it is hard to ensure that it is true in particular instances so as to solve the epistemological problem.

In the following sections, we begin with a review of four approaches of what causality might be. We spend most of our time on a counterfactual definition, mostly amounting to a recipe that is now widely used in statistics. We end with a discussion of the limitations of the recipe and how far it goes toward solving the epistemological and ontological problems.

4 The Humean and Neo-Humean Approach to Causation

4.1 Lawlike Generalities and the Humean Regularity Approach to Causation

Humean and neo-Humean approaches propose logical conditions that must hold for the constant conjunction of events to justify the inference that they have a cause–effect relationship. Specifically, Humeans have explored whether a cause must be sufficient for its effects, necessary for its effects, or something more complicated.

[15] Something different might also be needed. Hume himself dropped the requirement for contiguity in his 1748 rewrite of his 1738 work, and many philosophers would also drop his requirement for temporal precedence.

[16] Those new to this literature are presented with many statements about the need for lawfulness and unconditionality which seem to promise a recipe that will insure lawfulness. But the conditions that are presented always seem to fall short of the goal.

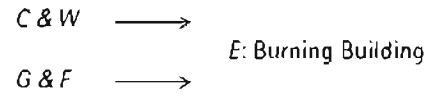

Fig. 10.1. Two sets of INUS conditions

The classic definition shared by Hume, John Stuart Mill, and many others was that "X is a cause of Y if and only if X is sufficient for Y." That is, the cause must always and invariably lead to the effect. Certainly an X that is sufficient for Y can be considered a cause, but what about the many putative causes that are not sufficient for their effect? Striking a match, for example, may be necessary for it to light, but it may not light unless there is enough oxygen in the atmosphere. Is striking a match never a cause of a match lighting? This leads to an alternative definition in which "X is a cause of Y if and only if X is necessary for Y." Under this definition, it is assumed that the cause (such as striking the match) must be present for the effect to occur, but it may not always be enough for the cause to actually occur (because there might not be enough oxygen). But how many causes are even necessary for their effects? If the match does not light after striking it, someone might use a blowtorch to light it so that striking the match is not even necessary for the match to ignite. Do we therefore assume that striking the match is never a cause of its lighting? Necessity and sufficiency seem unequal to the task of defining causation.[17]

These considerations led John Mackie to propose a set of conditions requiring that a cause be an insufficient [I] but necessary [N] part of a condition which is itself unnecessary [U] but exclusively sufficient [S] for the effect. These INUS conditions can be explained by an example. Consider two ways that the effect (E), which is a building burning down, might occur (see Figure 10.1). In one scenario the wiring might short-circuit and overheat, thus causing the wooden framing to burn. In another, a gasoline can might be next to a furnace that ignites and causes the gasoline can to explode. A number of factors here are INUS conditions for the building to burn down. The short circuit (C) and the wooden framing (W) together might cause the building to burn down, or the gasoline can (G) and the furnace (F) might cause the building to burn down. Thus, C and W together are exclusively sufficient [S] to burn the building down, and G and F together are exclusively sufficient [S] to burn the building down. Furthermore, the short circuit and wooden framing (C *and* W) are unnecessary [U], and the gasoline can and the furnace (G *and* F) are unnecessary [U] because the building could have burned down with just one or the other combination of factors. Finally, C, W, G, or F alone is insufficient [I] to burn the building down even though C is necessary [N] in conjunction with W (or vice versa) and G is necessary [N] in conjunction with F (or vice-versa). This formulation allows for the fact that no single cause is sufficient or necessary, but when experts say that a short circuit caused the

[17] And there are problems such as the following favorite of the philosophers. "If two bullets pierce a man's heart simultaneously, it is reasonable to suppose that each is an essential part of a distinct sufficient condition of the death, and that neither bullet is *ceteris paribus* necessary for the death, since in each case the other bullet is sufficient" (Sosa and Tooley 1993, 8–9)

fire they "are saying, in effect that the short-circuit (C) is a condition of this sort, that it occurred, that the other conditions (W) which, conjoined with it, form a sufficient condition were also present, and that no other sufficient condition (such as G *and* F) of the house's catching fire was present on this occasion" (Mackie 1965, 245; letters addded).

From the perspective of a practicing researcher, three lessons follow from the INUS conditions. First a putative cause such as C might not cause the effect E because G *and* F might be responsible. Hence, the burned-down building (E) will not always result from a short circuit (C) even though C could cause the building to burn down. Second, interactions among causes may be necessary for any one cause to be sufficient (C and W require each other and W and G require each other). Third, the relationship between any INUS cause and its effect might appear to be probabilistic because of the other INUS causes. In summary, the INUS conditions suggest the multiplicity of causal pathways and causes, the possibility of conjunctural causation (Ragin 1987), and the likelihood that social science relationships will appear probabilistic even if they are deterministic.[18]

A specific example might help to make these points clearer. Assume that the four INUS factors mentioned above, C, W, G, and F, occur independently of one another and that they are the only factors which cause fires in buildings. Further assume that short circuits (C) occur 10 percent of the time, wooden (W) frame buildings 50 percent of the time, furnaces (F) 90 percent of the time, and gasoline (G) cans near furnaces 10 percent of the time. Because these events are assumed independent of one another, it is easy to calculate that C and W occur 5 percent of the time and that G and F occur 9 percent of the time. (We simply multiply the probability of the two independent events.) All four conditions occur 0.45 percent of the time. (The product of all four percentages.) Thus, fires occur 13.55 percent of the time. This percentage includes the cases where the fire is the result of C and W (5 percent of the time) and where it is the result of G and F (9 percent of the time), and it adjusts downward for double-counting that occurs in the cases where all four INUS conditions occur together (0.45 percent of the time).

Now suppose an experimenter did not know about the role of wooden frame buildings or gasoline cans and furnaces and only looked at the relationship between fires and short circuits. A cross-tabulation of fires with the short circuit factor would yield Table 10.2. As assumed above, short circuits occur 10 percent of the time (see the third column total at the bottom of the table) and as calculated above, fires occur 13.55 percent of the time (see the third row total on the far right). The entries in the interior of the table are calculated in a similar way.[19]

Even though each case occurs because of a deterministic process—either a short circuit and a wooden frame building or a gasoline can and a furnace (or both)—this cross-tabulation suggests a probabilistic relationship between fires and short

[18] These points are made especially forcefully in Marini and Singer (1988).

[19] Thus, the entry for short circuits and fires comes from the cases where there are short circuits and wooden frame buildings (5 percent of the time) and where there are short circuits and no wooden frame buildings but there are gasoline cans and furnaces (5 percent times 9 percent).

Table 10.2. Fires by short circuits in hypothetical example (total percentages of each event)

	Not C—no short circuits	C—short circuits	Row totals
Not E—no fires	81.90	4.55	86.45
E—fires	8.10	5.45	13.55
Column totals	90.00	10.00	100.00

circuits. In 4.55 percent of the cases, short circuits occur but no fires result because the building was not wooden. In 8.10 percent of the cases, there are no short circuits, but a fire occurs because the gasoline can has been placed near the furnace. For this table, a standard measure of association, the Pearson correlation, between the effect and the cause is about .40 which is far short of the 1.0 required for a perfect (positive) relationship. If, however, the correct model is considered in which there are the required interaction effects, the relationship will produce a perfect fit.[20] Thus, a misspecification of a deterministic relationship can easily lead a researcher to think that there is a probabilistic relationship between the cause and effect.

INUS conditions reveal a lot about the complexities of causality, but as a definition of it, they turn out to be too weak—they do not rule out situations where there are common causes, and they do not exclude accidental regularities. The problem of common cause arises in a situation where, for example, lightning strikes (L) the wooden framing (W) and causes it to burn (E) while also causing a short in the circuitry (C). That is, $L \rightarrow E$ and $L \rightarrow C$ (where the arrow indicates causation). If lightning always causes a short in the circuitry, but the short never has anything to do with a fire in these situations because the lightning starts the fire directly through its heating of the wood, we will nevertheless always find that C and E are constantly conjoined through the action of the lightning, suggesting that the short circuit caused the fire even though the truth is that *lightning is the common cause of both.*[21] In some cases of common causes such as the rise in barometric pressure followed by the arrival of a storm, common sense tells us that the putative cause (the rise in barometric pressure) cannot be the real cause of the thunderstorm. But in the situation with the lightning, the fact that short circuits have the capacity to cause fires makes it less likely that we will realize that lightning is the common cause of both the short circuits and the fires. We might be better off in the case where the lightning split some of the wood framing of the house instead of causing a short circuit. In that case,

[20] If each variable is scored zero or one depending upon whether the effect or cause is present or absent, then a regression equation of the effect on the product (or interaction) of C and W, the product of G and F, and the product of C, W, G, and F will produce a multiple correlation of one indicating a perfect fit.

[21] It is also possible that the lightning's heating of the wood is (always or sometimes) insufficient to cause the fire (not $L \rightarrow E$), but its creation of a short circuit ($L \rightarrow C$) is (always or sometimes) sufficient for the fire ($C \rightarrow E$). In this case, the *lightning is the indirect cause of the fire* through its creation of the short circuit. That is, $L \rightarrow C \rightarrow E$.

we would probably reject the fantastic theory that split wood caused the fire because split wood does not have the capacity to start a fire, but the Humean approach would be equally confused by both situations because it could not appeal, within the ambit of its understanding, to causal capacities. For a Humean, the constant conjunction of split wood and fires suggests causation as much as the constant conjunction of short circuits and fires. Indeed, the constant conjunction of storks and babies would be treated as probative of a causal connection.

Attempts to fix up these conditions usually focus on trying to require "lawlike" statements that are unconditionally true, not just accidentally true. Since it is not unconditionally true that splitting wood causes fires, the presumption is that some such conditions can be found to rule out this explanation. Unfortunately, no set of conditions seem to be successful.[22] Although the regularity approach identifies a necessary condition for describing causation, it basically fails because association is not causation and there is no reason why purely logical restrictions on lawlike statements should be sufficient to characterize causal relationships. Part of the problem is that there are many different types of causal laws and they do not fit any particular patterns. For example, one restriction that has been proposed to ensure lawfulness is that lawlike statements should either not refer to particular situations or they should be derivable from laws that do not refer to particular situations. This would mean that Kepler's first "law" about all planets moving in elliptical orbits around the sun (a highly specific situation!) was not a causal law before Newton's laws were discovered, but it was a causal law after it was shown that it could be derived from Newton's laws. But Kepler's laws were always considered causal laws, and there seems to be no reason to rest their lawfulness on Newton's laws. Furthermore, by this standard, almost all social science and natural science laws (e.g. plate tectonics) are about particular situations. In short, logical restrictions on the form of laws do not seem sufficient to characterize causality.

4.2 The Asymmetry of Causation

The regularity approach also fails because it does not provide an explanation for the asymmetry of causation. Causes should cause their effects, but INUS conditions are almost always symmetrical such that if C is an INUS cause of E, then E is also an INUS cause of C. It is almost always possible to turn around an INUS condition so that an effect is an INUS for its cause.[23] One of the most famous examples of this problem involves a flagpole, the elevation of the sun, and the flagpole's shadow. The

[22] For some representative discussions of the problems see Harre and Madden (1975, ch. 2); Salmon (1990, chs. 1–2); Hausman (1998, ch. 3). Salmon (1990, 15) notes that "Lawfulness, modal import [what is necessary, possible, or impossible], and support of counterfactuals seems to have a common extension: statements either possess all three or lack all three. But it is extraordinarily difficult to find criteria to separate those statements that do from those that do not."

[23] Papineau (1985, 279) provides a demonstration of the symmetry of INUS conditions, and he goes on to suggest a condition for the asymmetry of causation that does not rely on the temporal relationship between causes and effects.

law that light travels in straight lines implies that there is a relationship between the height of the flagpole, the length of its shadow, and the angle of elevation of the sun. When the sun rises, the shadow is long, at midday it is short, and at sunset it is long again. Intuition about causality suggests that the length of the shadow is caused by the height of the flagpole and the elevation of the sun. But, using INUS conditions, we can just as well say that the elevation of the sun is caused by the height of the flagpole and the length of the shadow. There is simply nothing in the conditions that precludes this fantastic possibility.

The only feature of the Humean approach that provides for asymmetry is temporal precedence. If changes in the elevation of the sun precede corresponding changes in the length of the shadow, then we can say that the elevation of the sun causes the length of the shadow. And if changes in the height of the flagpole precede corresponding changes in the length of the shadow, we can say that the height of the flagpole causes the length of the shadow. But many philosophers reject making temporal precedence the determinant of causal asymmetry because it precludes the possibility of *explaining* the direction of time by causal asymmetry and it precludes the possibility of backwards causation. From a practical perspective, it also requires careful measures of timing that may be difficult in a particular situation.

4.3 Summary

This discussion reveals two basic aspects of the causal relation. One is a symmetrical form of association between cause and effect and the other is an asymmetrical relation in which causes produce effects but not the reverse. The Humean regularity approach, in the form of INUS conditions, provides a necessary condition for the existence of the symmetrical relationship,[24] but it does not rule out situations such as common cause and accidental regularities where there is no causal relationship at all. From a methodological standpoint, it can easily lead researchers to presume that all they need to do is to find associations, and it also leads to an underemphasis on the rest of the requirement for a "lawlike" or "unconditional" relationship because it does not operationally define what that would really mean. A great deal of what passes for causal modeling suffers from these defects (Freedman 1987; 1991; 1997; 1999).

The Humean approach does even less well with the asymmetrical feature of the causal relationship because it provides no way to determine asymmetry except temporal precedence. Yet there are many other aspects of the causal relation that seem more fundamental than temporal precedence. Causes not only typically precede their

[24] Probabilistic causes do not necessarily satisfy INUS conditions because an INUS factor might only sometimes produce an effect. Thus, the short circuit and the wooden frame of the house might only sometimes lead to a conflagration in which the house is burned down. Introducing probabilistic causes would add still another layer of complexity to our discussion which would only provide more reasons to doubt the Humean regularity approach.

effects, but they also can be used to explain effects or to manipulate effects while effects cannot be used to explain causes or to manipulate them.[25]

Effects also depend upon causes, but causes do not depend upon effects. Thus, if a cause does not occur, then the effect will not occur because effects depend on their causes. The counterfactual, "if the cause did not occur, then the effect would not occur," is true. However, if the effect does not occur, then the cause might still occur because causes can happen without leading to a specific effect if other features of the situation are not propitious for the effect. The counterfactual, "if the effect did not occur, then the cause would not occur," is not necessarily true. For example, where a short circuit causes a wooden frame building to burn down, if the short circuit does not occur, then the building will not burn down. But if the building does not burn down, it is still possible that the short circuit occurred but its capacity for causing fires was neutralized because the building was made of brick. This dependence of effects on causes suggests that an alternative definition of causation might be based upon a proper understanding of counterfactuals.

5 Counterfactual Definition of Causation

In a book *On the Theory and Method of History* published in 1902, Eduard Meyer claimed that it was an "unanswerable and so an idle question" whether the course of history would have been different if Bismarck, then Chancellor of Prussia, had not decided to go to war in 1866. By some accounts, the Austro-Prussian-Italian War of 1866 paved the way for German and Italian unification (see Wawro 1996). In reviewing Meyer's book in 1906, Max Weber agreed that "from the strict 'determinist' point of view" finding out what would have happened if Bismarck had not gone to war "was 'impossible' given the 'determinants' which were in fact present." But he went on to say that "And yet, for all that, it is far from being 'idle' to raise the question what might have happened, if, for example, Bismarck had not decided for war. For it is precisely this question which touches on the decisive element in the historical construction of reality: the causal significance which is properly attributed to this individual decision within the totality of infinitely numerous 'factors' (all of which must be just as they are and not otherwise) if precisely this consequence is to result, and the appropriate position which the decision is to occupy in the historical account" (Weber 1978, 111). Weber's review is an early discussion of the importance of counterfactuals for understanding history and making causal inferences. He argues forcefully that if "history is to raise itself above the level of a mere chronicle of noteworthy events and

[25] Hausman (1998, 1) also catalogs other aspects of the asymmetry between causes and effects.

personalities, it can only do so by posing just such questions" as the counterfactual in which Bismarck did not decide for war.[26]

5.1 Lewis's Counterfactual Approach to Causation

The philosopher David Lewis (1973*b*) has proposed the most elaborately worked out theory of how causality is related to counterfactuals.[27] His approach requires the truth of two statements regarding two distinct events *X* and *Y*. Lewis starts from the presumption that *X* and *Y* have occurred so that the "counterfactual" statement,[28] "If *X* were to occur, then *Y* would occur," is true. The truth of this statement is Lewis's first condition for a causal relationship. Then he considers the truth of a second counterfactual:[29] "If *X* were not to occur, then *Y* would not occur either." If this is true as well, then he says that *X* causes *Y*. If, for example, Bismarck decided for war in 1866 and, as some historians argue, German unification followed because of his decision, then we must ask: "If Bismarck had not decided for war, would Germany have remained divided?" The heart of Lewis's approach is the set of requirements, described below, that he lays down for the truth of this kind of counterfactual.

Lewis's theory has a number of virtues. It deals directly with singular causal events, and it does not require the examination of a large number of instances of *X* and *Y*. At one point in the philosophical debate about causation, it was believed that the individual cases such as "the hammer blow caused the glass to break" or "the assassination of Archduke Ferdinand caused the First World War" could not be analyzed alone because these cases had to be subsumed under a general law ("hammer blows cause glass to break") derived from multiple cases plus some particular facts of the situation in order to meet the requirement for a "lawlike" relationship. The counterfactual approach, however, starts with singular events and proposes that causation can be established without an appeal to a set of similar events and general

[26] I am indebted to Richard Swedberg for pointing me towards Weber's extraordinary discussion.

[27] Lewis finds some support for his theory in the work of David Hume. In a famous change of course in a short passage in his *Enquiry Concerning Human Understanding* (1748), Hume first summarized his regularity approach to causation by saying that "we may define a cause to be an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second," and then he changed to a completely different approach to causation by adding "Or in other words, where if the first object had not been, the second had never existed" (146). As many commentators have noted, these were indeed other words, implying an entirely different notion of causation. The first approach equates causality with the constant conjunction of putative causes and effects across similar circumstances. The second, which is a counterfactual approach, relies upon what would happen in a world where the cause did not occur.

[28] Lewis considers statements like this as part of his theory of counterfactuals by simply assuming that statements in the subjunctive mood with true premises and true conclusions are true. As noted earlier, most theories of counterfactuals have been extended to include statements with true premises by assuming, quite reasonably, that they are true if their conclusion is true and false otherwise.

[29] This is a simplified version of Lewis's theory based upon Lewis (1973*a*; 1973*b*; 1986) and Hausman (1998, ch. 6).

laws regarding them.[30] The possibility of analyzing singular causal events is important for all researchers, but especially for those doing case studies who want to be able to say something about the consequences of Stalin succeeding Lenin as head of the Soviet Union or the impact of the butterfly ballot on the 2000 US election.

The counterfactual approach also deals directly with the issue of *X*'s causal "efficacy" with respect to *Y* by considering what would happen if *X* did not occur. The problem with the theory is the difficulty of determining the truth or falsity of the counterfactual "If *X* were not to occur, then *Y* would not occur either." The statement cannot be evaluated in the real world because X actually occurs so that the premise is false, and there is no evidence about what would happen if X did not occur. It only makes sense to evaluate the counterfactual in a world in which the premise is true. Lewis's approach to this problem is to consider whether the statement is true in the closest possible world to the actual world where *X* does not occur. Thus, if *X* is a hammer blow and *Y* is a glass breaking, then the closest possible world is one in which everything else is the same except that the hammer blow does not occur. If in this world, the glass does not break, then the counterfactual is true, and the hammer blow (*X*) causes the glass to break (*Y*). The obvious problem with this approach is identifying the closest possible world. If *X* is the assassination of Archduke Ferdinand and *Y* is the First World War, is it true that the First World War would not have occurred in the closest possible world where the bullet shot by the terrorist Gavrilo Princip did not hit the Archduke? Or would some other incident have inevitably precipitated the First World War? And, to add to the difficulty, would this "First World War" be the same as the one that happened in our world?

Lewis's approach substitutes the riddle of determining the similarity of possible worlds for the neo-Humean's problem of determining lawlike relationships. To solve these problems, both approaches must be able to identify similar causes and similar effects. The Humean approach must identify them across various situations in the real world. This aspect of the Humean approach is closely related to John Stuart Mill's "Method of Concomitant Variation" which he described as follows: "Whatever phenomenon varies in any manner, whenever another phenomenon varies in some similar manner, is either a cause or an effect of that phenomenon, or is connected to it through some fact of causation" (Mill 1888, 287).[31] Lewis's theory must also identify similar causes and similar effects in the real world in which the cause does occur and in the many possible worlds in which the cause does not occur. This approach is

[30] In fact, many authors now believe that general causation (involving lawlike generalizations) can only be understood in terms of singular causation: "general causation is a generalisation of singular causation. Smoking causes cancer iff (if and only if) smokers' cancers are generally caused by their smoking" (Mellors 1995, 6–7). See also Sosa and Tooley (1993). More generally, whereas explanation was once thought virtually to supersede the need for causal statements, many philosophers now believe that a correct analysis of causality will provide a basis for suitable explanations (see Salmon 1990).

[31] The Humean approach also has affinities with Mill's Method of Agreement which he described as follows: "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon" (Mill 1888, 280).

closely related to Mill's "Method of Difference" in which: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon" (Mill 1888, 280).[32]

In addition to identifying similar causes and similar effects, the Humean approach must determine if the conjunction of these similar causes and effects is accidental or lawlike. This task requires understanding what is happening in each situation and comparing the similarities and differences across situations. Lewis's approach must identify the possible world where the cause does not occur that is most similar to the real world. This undertaking requires understanding the facts of the real world and the laws that are operating in it. Consequently, assessing the similarity of a possible world to our own world requires understanding the lawlike regularities that govern our world.[33] It seems as if Lewis has simply substituted one difficult task, that of identifying the most similar world for the job of establishing lawfulness.

5.2 The Virtues of the Counterfactual Definition of Causation

Lewis *has* substituted one difficult problem for another, but the reformulation of the problem has a number of benefits. The counterfactual approach provides new insights into what is required to establish causal connection between causes and effects. The counterfactual approach makes it clear that establishing causation does not require observing the universal conjunction of a cause and an effect.[34] One observation of a cause followed by an effect is sufficient for establishing causation if it can be shown that in a most similar world without the cause, the effect does not occur. The counterfactual approach proposes that causation can be demonstrated by simply finding a most similar world in which the absence of the cause leads to the absence of the effect.

Lewis's theory provides us with a way to think about the causal impact of singular events such as the badly designed butterfly ballot in Palm Beach County, Florida that led some voters in the 2000 presidential election to complain that they mistakenly voted for Reform Party candidate Patrick Buchanan when they meant to vote for Democrat Al Gore. The ballot can be said to be causally associated with these mistakes

[32] Mill goes on to note that the Method of Difference is "a method of artificial experiment" (281). Notice that for both the Method of Concomitant Variation and the Method of Difference, Mill emphasizes the association between cause and effect and not the identification of which event is the cause and which is the effect. Mill's methods are designed to detect the symmetric aspect of causality but not its asymmetric aspect.

[33] Nelson Goodman makes this point in a 1947 article on counterfactuals, and James Fearon (1991), in a masterful exposition of the counterfactual approach to research, discusses its implications for counterfactual thought experiments in political science. Also see Tetlock and Belkin (1996).

[34] G. H. von Wright notes that the counterfactual conception of causality shows that the hallmark of a lawlike connection is "*necessity and not universality*" (von Wright 1971, 22).

if in the closest possible world in which the butterfly ballot was not used, the vote for Buchanan was lower than in the real world. Ideally this closest possible world would be a parallel universe in which the same people received a different ballot, but this, of course, is impossible. The next-best thing is a situation where similar people employed a different ballot. In fact, the butterfly ballot was only used for election day voters in Palm Beach County. It was not used by absentee voters. Consequently, the results for the absentee voting can be considered a surrogate for the closest possible world in which the butterfly ballot was not used, and in this absentee voting world, voting for Buchanan was dramatically lower, suggesting that at least 2000 people who preferred Gore—more than enough to give the election to Gore—mistakenly voted for Buchanan on the butterfly ballot.

The difficult question, of course, is whether the absentee voting world can be considered a good enough surrogate for the closest possible world in which the butterfly ballot was not used.[35] The counterfactual approach does not provide us with a clear sense of how to make that judgment.[36] But the framework does suggest that we should consider the similarity of the election day world and the absentee voter world. To do this, we can ask whether election day voters are different in some significant ways from absentee voters, and this question can be answered by considering information on their characteristics and experiences. In summary, the counterfactual perspective allows for analyzing causation in singular instances, and it emphasizes comparison, which seems difficult but possible, rather than the recondite and apparently fruitless investigation of the lawfulness of statements such as "All ballots that place candidate names and punch-holes in confusing arrangements will lead to mistakes in casting votes."

5.3 Controlled Experiments and Closest Possible Worlds

The difficulties with the counterfactual definition are identifying the characteristics of the closest possible world in which the putative cause does not occur and finding an empirical surrogate for this world. For the butterfly ballot, sheer luck led a team of researchers to discover that the absentee ballot did not have the problematic features of the butterfly ballot.[37] But how can we find surrogates in other circumstances?

[35] For an argument that the absentee votes are an excellent surrogate, see Wand et al. (1991).

[36] In his book on counterfactuals, Lewis only claims that similarity judgments are possible, but he does not provide any guidance on how to make them. He admits that his notion is vague, but he claims it is not ill-understood. "But comparative similarity is not ill-understood. It is vague—very vague—in a well-understood way. Therefore it is just the sort of primitive that we must use to give a correct analysis of something that is itself undeniably vague" (Lewis 1973*a*, 91). In later work Lewis (1979; 1986) formulates some rules for similarity judgments, but they do not seem very useful to us and to others (Bennett 1988).

[37] For the story of how the differences between the election day and absentee ballot were discovered, see Brady et al. (2001).

One answer is controlled experiments. Experimenters can create mini-closest-possible worlds by finding two or more situations and assigning putative causes (called "treatments") to some situations but not to others (which get the "control"). If in those cases where the cause C occurs, the effect E occurs, then the first requirement of the counterfactual definition is met: When C occurs, then E occurs. Now, if the situations which receive the control are not different in any significant ways from those that get the treatment, then they can be considered surrogates for the closest possible world in which the cause does not occur. If in these situations where the cause C does not occur, the effect E does not occur either, then the second requirement of the counterfactual definition is confirmed: In the closest possible world where C does not occur, then E does not occur. The crucial part of this argument is that the control situation, in which the cause does not occur, must be a good surrogate for the closest possible world to the treatment.

Two experimental methods have been devised for ensuring closeness between the treatment and control situations. One is classical experimentation in which as many circumstances as possible are physically controlled so that the only significant difference between the treatment and the control is the cause. In a chemical experiment, for example, one beaker holds two chemicals and a substance that might be a catalyst and another beaker of the same type, in the same location, at the same temperature, and so forth contains just the two chemicals in the same proportions without the suspected catalyst. If the reaction occurs only in the first beaker, it is attributed to the catalyst. The second method is random assignment of treatments to situations so that there are no reasons to suspect that the entities that get the treatment are any different, on average, from those that do not. We discuss this approach in detail below.

5.4 Problems with the Counterfactual Definition[38]

Although the counterfactual definition of causation leads to substantial insights about causation, it also leads to two significant problems. Using the counterfactual definition as it has been described so far, the direction of causation cannot be established, and two effects of a common cause can be mistaken for cause and effect. Consider, for example, an experiment as described above. In that case, in the treatment group, when C occurs, E occurs, and when E occurs, C occurs. Similarly, in the control group, when C does not occur, then E does not occur, and when E does not occur, then C does not occur. In fact, there is perfect observational symmetry between cause and effect which means that the counterfactual definition of causation as described so far implies that C causes E *and* that E causes C. The same problem arises with two effects of a common cause because of the perfect symmetry in the situation. Consider, for example, a rise in the mercury in a barometer and thunderstorms. Each is an effect

[38] This section relies heavily upon Hausman (1998, especially chs. 4–7) and Lewis (1973*b*).

of high pressure systems, but the counterfactual definition would consider them to be causes of one another.[39]

These problems bedevil Humean and counterfactual approaches. If we accept these approaches in their simplest forms, we must live with a seriously incomplete theory of causation that cannot distinguish causes from effects and that cannot distinguish two effects of a common cause from real cause and effect. That is, although the counterfactual approach can tell whether two factors A and B are causally connected[40] in some way, it cannot tell whether A causes B, B causes A, or A and B are the effects of a common cause (sometimes called spurious correlation). The reason for this is that the truth of the two counterfactual conditions described so far amounts to a particular pattern of the cross-tabulation of the two factors A and B. In the simplest case where the columns are the absence or presence of the first factor (A) and the rows are the absence or the presence of the second factor (B), then the same diagonal pattern is observed for situations where A causes B or B causes A, or for A and B being the effects of a common cause. In all three cases, we either observe the presence of both factors or their absence. It is impossible from this kind of symmetrical information, which amounts to correlational data, to detect causal asymmetry or spurious correlation. The counterfactual approach as elucidated so far, like the Humean regularity approach, only describes a necessary condition, the existence of a causal connection between A and B, for us to say that A causes B.

Requiring temporal precedence can solve the problem of causal direction by simply choosing the phenomenon that occurs first as the cause, but it cannot solve the problem of common cause because it would lead to the ridiculous conclusion that since the mercury rises in barometers before storms, this upward movement in the mercury must cause thunderstorms. For this and other reasons, David Lewis rejects using temporal precedence to determine the direction of causality. Instead, he claims that when C causes E but not the reverse "then it should be possible to claim the falsity of the counterfactual 'If E did not occur, then C would not occur.'" This counterfactual is different from "if C occurs then E occurs" and from "if C does not occur then E does not occur" which, as we have already mentioned, Lewis believes must both be true when C causes E. The required falsity of "If E did not occur, then C would not occur" adds a third condition for causality. This condition amounts to finding situations in which C occurs but E does not—typically because there is some other condition that must occur for C to produce E. Rather than explore this strategy, we describe a much better way of establishing causal priority in the next section.

[39] Thus, if barometric pressure rises, thunderstorms occur and vice versa. Furthermore, if barometric pressure does not rise, then thunderstorms do not occur and vice versa. Thus, by the counterfactual definition, each is the cause of the other. (To simplify matters, we have ignored the fact that there is not a perfectly deterministic relationship between high pressure systems and thunderstorms.)

[40] As implied by this paragraph, there is a causal connection between A and B when either A causes B, B causes A, or A and B are the effects of a common cause. See Hausman (1998, 55–63).

6 Experimentation and the Manipulation Approach to Causation

In an experiment, there is a readily available piece of information that we have overlooked so far because it is not mentioned in the counterfactual approach. The factor that has been manipulated can determine the direction of causality and help to rule out spurious correlation. The manipulated factor must be the cause.[41] It is hard to exaggerate the importance of this insight. Although philosophers are uncomfortable with manipulation and agency approaches to causality because they put people (as the manipulators) at the center of our understanding of causality, there can be little doubt about the power of manipulation for determining causality. Agency and manipulation approaches to causation (Gasking 1955; von Wright 1974; Menzies and Price 1993) elevate this insight into their definition of causation. For Gasking "the notion of causation is essentially connected with our manipulative techniques for producing results" (1955, 483), and for Menzies and Price "events are causally related just in case the situation involving them possesses intrinsic features that *either* support a means-end relation between the events as is, *or* are identical with (or closely similar to) those of another situation involving an analogous pair of means-end related events" (1993, 197). These approaches focus on establishing the direction of causation, but Gasking's metaphor of causation as "recipes" also suggests an approach towards establishing the symmetric, regularity aspect of causation. Causation exists when there is a recipe that regularly produces effects from causes.

Perhaps our ontological definitions of causality should not employ the concept of agency because most of the causes and effects in the universe go their merry way without human intervention, and even our epistemological methods often discover causes, as with Newtonian mechanics or astrophysics, where human manipulation is impossible. Yet our epistemological methods cannot do without agency because human manipulation appears to be the best way to identify causes, and many researchers and methodologists have fastened upon experimental interventions as the way to pin down causation. These authors typically eschew ontological aims and emphasize epistemological goals. After explicitly rejecting ontological objectives, for example, Herbert Simon proceeds to base his initial definition of causality on experimental systems because "in scientific literature the word 'cause' most often occurs in connection with some explicit or implicit notion of an experimenter's intervention in a system" (Simon 1952, 518). When full experimental control is not possible, Thomas Cook and Donald T. Campbell recommend "quasi-experimentation," in which "an abrupt intervention at a known time" in a treatment group makes it possible to compare

[41] It might be more correct to say that the cause is buried somewhere among those things that were manipulated or that are associated with the manipulation. It is not always easy, however, to know what was manipulated as in the famous Hawthorne experiments in which the experimenters thought the treatment was reducing the lighting for workers but the workers apparently thought of the treatment as being treated differently from all other workers. Part of the work required for good causal inference is clearly describing what was manipulated and unpacking it to see what feature caused the effect.

the impacts of the treatment over time or across groups (Cook and Campbell 1986, 149). The success of quasi-experimentation depends upon "a world of probabilistic multivariate causal agency in which some manipulable events dependably cause other things to change" (150). John Stuart Mill suggests that the study of phenomena which "we can, by our voluntary agency, modify or control" makes it possible to satisfy the requirements of the Method of Difference ("a method of artificial experiment") even though "by the spontaneous operations of nature those requisitions are seldom fulfilled" (Mill 1888, 281, 282). Sobel champions a manipulation model because it "provides a framework in which the nonexperimental worker can think more clearly about the types of conditions that need to be satisfied in order to make inferences" (Sobel 1995, 32). David Cox claims that quasi-experimentation "with its interventionist emphasis seems to capture a deeper notion" (Cox 1992, 297) of causality than the regularity approach.

As we shall see, there are those who dissent from this perspective, but even they acknowledge that there is "wide agreement that the idea of causation as consequential manipulation is stronger or 'deeper' than that of causation as robust dependence" (Goldthorpe 2001, 5). This account of causality is especially compelling if the manipulation approach and the counterfactual approach are conflated, as they often are, and viewed as one approach. Philosophers seldom combine them into one perspective, but all the methodological writers cited above (Simon, Cook and Campbell, Mill, Sobel, and Cox) conflate them because they draw upon controlled experiments, which combine intervention and control, for their understanding of causality. Through interventions, experiments manipulate one (or more) factor which simplifies the job of establishing causal priority by appeal to the manipulation approach to causation. Through laboratory controls or statistical randomization, experiments also create closest possible worlds that simplify the job of eliminating confounding explanations by appeal to the counterfactual approach to causation.

The combination of intervention and control in experiments makes them especially effective ways to identify causal relationships. If experiments only furnished closest possible worlds, then the direction of causation would be indeterminate without additional information. If experiments only manipulated factors, then accidental correlation would be a serious threat to valid inferences about causality. Both features of experiments do substantial work.

Any approach to determining causation in nonexperimental contexts that tries to achieve the same success as experiments must recognize both these features. The methodologists cited above conflate them, and the psychological literature on counterfactual thinking cited at the beginning of this chapter shows that our natural inclination as human beings is to conflate them. When considering alternative possibilities, people typically consider nearby worlds in which individual agency figures prominently. When asked to consider what could have happened differently in a vignette involving a drunken driver and a new route home from work, subjects focus on having taken the new route home instead of on the factors that led to drunken driving. They choose a cause and a closest possible world in which *their* agency matters. But there is no reason why the counterfactual approach and the manipulation

approach should be combined in this way. The counterfactual approach to causation emphasizes possible worlds without considering human agency and the manipulation approach to causation emphasizes human agency without saying anything about possible worlds. Experiments derive their strength from combining both theoretical perspectives, but it is all too easy to overlook one of these two elements in generalizing from experimental to observational studies.[12]

As we shall see in a later section, the best-known statistical theory of causality emphasizes the counterfactual aspects of experiments without giving equal attention to their manipulative aspects. Consequently, when the requirements for causal inference are transferred from the experimental setting to the observational setting, those features of experiments that rest upon manipulation tend to get underplayed.

7 Pre-emption and the Mechanism Approach to Causation

7.1 Pre-emption

Experimentation's amalgamation of the lessons of counterfactual and manipulation approaches to causation produces a powerful technique for identifying the effects of manipulated causes. Yet, in addition to the practical problems of implementing the recipe correctly, the experimental approach does not deal well with two related problems. It does not solve the problem of causal pre-emption which occurs when one cause acts just before and pre-empts another, and it does not so much explain the causes of events as it demonstrates the effects of manipulated causes. In both cases, the experimentalists' focus on the impacts of manipulations in the laboratory instead of on the causes of events in the world, leads to a failure to explain important phenomena, especially those phenomena which cannot be easily manipulated or isolated.

The problem of pre-emption illustrates this point. The following example of pre-emption is often mentioned in the philosophical literature. A man takes a trek across a desert. His enemy puts a hole in his water can. Another enemy, not knowing the action of the first, puts poison in his water. Manipulations have certainly occurred, and the man dies on the trip. The enemy who punctured the water can thinks that she

[12] Some physical experiments actually derive most of their strength by employing such powerful manipulations that no controls are needed. At the detonation of the first atom bomb, no one doubted that the explosion was the result of nuclear fission and not some other uncontrolled factor. Similarly, in what might be an apocryphal story, it is said that a Harvard professor who was an expert on criminology once lectured to a class about how all social science evidence suggested that rehabilitating criminals simply did not work. A Chinese student raised his hand and politely disagreed by saying that during the Cultural Revolution, he had observed cases where criminals had been rehabilitated. Once again, a powerful manipulation may need no controls.

caused the man to die, and the enemy who added the poison thinks that he caused the man to die. In fact, the water dripping out of the can pre-empted the poisoning so that the poisoner is wrong. This situation poses problems for the counterfactual approach because one of the basic counterfactual conditions required to establish that the hole in the water can caused the death of the man, namely the truth of the counterfactual "if the hole had not been put in the water can, the man would not have died," is false even though the man did in fact die of thirst. The problem is that the man would have died of poisoning if the hole in the water can had not pre-empted that cause, and the "back-up" possibility of dying by poisoning falsifies the counterfactual.

The pre-emption problem is a serious one, and it can lead to mistakes even in well-designed experiments. Presumably the closest possible world to the one in which the water can has been punctured is one in which the poison has been put in the water can as well. Therefore, even a carefully designed experiment will conclude that the puncturing of the can did not kill the man crossing the desert because the unfortunate subject in the control condition would die (from poisoning) just as the subject in the treatment would die (from the hole in the water can). The experiment alone would not tell us how the man died. A similar problem could arise in medical experiments. Arsenic was once used to cure venereal disease, and it is easy to imagine an experiment in which doses of arsenic "cure" venereal disease but kill the patient while the members of the control group without the arsenic die of venereal disease at the same rate. If the experiment simply looked at the mortality rates of the patients, it would conclude that arsenic had no medicinal value because the same number of people died in the two conditions.

In both these instances, the experimental method focuses on the effects of causes and not on explaining effects by adducing causes. Instead of asking why the man died in his trek across the desert, the experimental approach asks what happens when a hole is put in the man's canteen and everything else remains the same. The method concludes that the hole had no effect. Instead of asking what caused the death of the patients with venereal disease, the experimental method asks whether giving arsenic to those with venereal disease had any net impact on mortality rates. It concludes that it did not. In short, experimental methods do not try to explain events in the world so much as they try to show what would happen if some cause were manipulated. This does not mean that experimental methods are not useful for explaining what happens in the world, but it does mean that they sometimes miss the mark.

7.2 Mechanisms, Capacities, and the Pairing Problem

The pre-emption problem is a vivid example of a more general problem with the Humean account that requires a solution. The general problem is that constant conjunction of events is not enough to "pair-up" particular events even when pre-emption is not present. Even if we know that holes in water cans generally spell trouble for desert travelers, we still have the problem of linking a particular hole in a water can with a particular death of a traveler. Douglas Ehring notes that:

> Typically, certain spatial and temporal relations, such as spatial/temporal contiguity, are invoked to do this job. [That is, the hole in the water can used by the traveler is obviously the one that caused his death because it is spatially and temporally contiguous to him.] These singularist relations are intended to solve the residual problem of causally pairing particular events, a problem left over by the generalist core of the Humean account. (Ehring 1997, 18)

Counterfactual approaches, because they can explain singular causal events, do not suffer so acutely from this "pairing" problem, but the pre-emption problem shows that remnants of the difficulty remain even in counterfactual accounts (Ehring 1997, ch. 1). In both the desert traveler and arsenic examples, the counterfactual account cannot get at the proper pairing of causes and effects because there are two redundant causes to be paired with the same effects. Something more is needed.

The solution in both these cases seems obvious, but it does not follow from the neo-Humean, counterfactual, or manipulation definitions of causality. The solution is to inquire more deeply into what is happening in each situation in order to describe the capacities and mechanisms that are operating. An autopsy of the desert traveler would show that the person died of thirst, and an examination of the water can would show that the water would have run out before the poisoned water could be imbibed. An autopsy of those given arsenic would show that the signs of venereal disease were arrested while other medical problems, associated with arsenic poisoning, were present. Further work might even show that lower doses of arsenic cure the disease without causing death. In both these cases, deeper inquires into the mechanism by which the causes and effects are linked would produce better causal stories.

But what does it mean to explicate mechanisms and capacities?[13] "Mechanisms" we are told by Machamber, Darden, and Craver (2000, 3) "are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." The crucial terms in this definition are "entities and activities" which suggest that mechanisms have pieces. Glennan (1996, 52) calls them "parts," and he requires that it should be possible "to take the part out of the mechanism and consider its properties in another context." Entities, or parts, are organized to produce change. For Glennan (52), this change should be produced by "the interaction of a number of parts according to direct causal laws." The biological sciences abound with mechanisms of this sort such as the method of DNA replication, chemical transmission at synapses, and protein synthesis. But there are many mechanisms in the social sciences as well including markets with their methods of transmitting price information and bringing buyers and sellers together, electoral systems with their routines for bringing candidates and voters together in a collective decision-making process, the diffusion of innovation through

[13] These approaches are not the same, and those who favor one often reject the other (see, e.g., Cartwright 1989 on capacities and Machamer, Darden, and Craver 2000 on mechanisms). But both emphasize "causal powers" (Harre and Madden 1975, ch. 5) instead of mere regularity or counterfactual association. We focus on mechanisms because we believe that they are a somewhat better way to think about causal powers, but in keeping with our pragmatic approach, we find much that is useful in "capacity" approaches

social networks, the two-step model of communication flow, weak ties in social networks, dissonance reduction, reference groups, arms races, balance of power, etc. (Hedstrom and Swedberg 1998). As these examples demonstrate, mechanisms are not exclusively mechanical, and their activating principles can range from physical and chemical processes to psychological and social processes. They must be composed of appropriately located, structured, and oriented entities which involve activities that have temporal order and duration, and "an activity is usually designated by a verb or verb form (participles, gerundives, etc.)" (Machamber, Darden, and Craver 2000, 4) which takes us back to the work of Lakoff and Johnson (1999) who identified a "Causation Is Forced Movement metaphor."

Mechanisms provide another way to think about causation. Glennan argues that "two events are causally connected when and only when there is a mechanism connecting them" and "the necessity that distinguishes connections from accidental conjunctions is to be understood as deriving from a underlying mechanism" which can be empirically investigated (64). These mechanisms, in turn, are explained by causal laws, but there is nothing circular in this because these causal laws refer to how the *parts* of the mechanism are connected. The operation of these parts, in turn, can be explained by lower-level mechanisms. Eventually the process gets to a bedrock of fundamental physical laws which Glennan concedes "cannot be explained by the mechanical theory" (65).

Consider explaining social phenomena by examining their mechanisms. Duverger's law, for example, is the observed tendency for just two parties in simple plurality single-member district elections systems (such as the United States). The entities in the mechanisms behind Duverger's law are voters and political parties. These entities face a particular electoral rule (single-district plurality voting) which causes two activities. One is that voters often vote strategically by choosing a candidate other than their most liked because they want to avoid throwing their vote away on a candidate who has no chance of winning and because they want to forestall the election of their least wanted alternative. The other activity is that political parties often decide not to run candidates when there are already two parties in a district because they anticipate that voters will spurn their third party effort.

These mechanisms underlying Duverger's law suggest other things that can be observed beyond the regularity of two-party systems being associated with single-member plurality-vote electoral systems that led to the law in the first place. People's votes should exhibit certain patterns and third parties should exhibit certain behaviors. And a careful examination of the mechanism suggests that in some federal systems that use simple plurality single-member district elections we might have more than two parties, seemingly contrary to Duverger's law. Typically, however, there are just two parties in each province or state, but these parties may differ from one state to another, thus giving the impression, at the national level, of a multiparty system even though Duverger's law holds in each electoral district.[44]

[44] This radically simplifies the literature on Duverger's law (see Cox 1997 for more details).

Or consider meterological[45] and physical phenomena. Thunderstorms are not merely the result of cold fronts hitting warm air or being located near mountains; they are the results of parcels of air rising and falling in the atmosphere subject to thermodynamic processes which cause warm humid air to rise, to cool, and to produce condensed water vapor. Among other things, this mechanism helps to explain why thunderstorms are more frequent in areas, such as Denver, Colorado, near mountains because the mountains cause these processes to occur—without the need for a cold air front. Similarly, Boyle's law is not merely a regularity between pressure and volume; it is the result of gas molecules moving within a container and exerting force when they hit the walls of the container. This mechanism for Boyle's law also helps to explain why temperature affects the relationship between the pressure and volume of a gas. When the temperature increases, the molecules move faster and exert more force on the container walls.

Mechanisms like these are midway between general laws on the one hand and specific descriptions on the other hand, and activities can be thought of as causes which are not related to lawlike generalities.[46] Mechanisms typically explicate observed regularities in terms of lower-level processes, and the mechanisms vary from field to field and from time to time. Moreover, these mechanisms "bottom out" relatively quickly—molecular biologists do not seek quantum mechanical explanations and social scientists do not seek chemical explanations of the phenomena they study.

When an unexplained phenomenon is encountered in a science, "Scientists in the field often recognize whether there are known types of entities and activities that can possibly accomplish the hypothesized changes and whether there is empirical evidence that a possible schemata is plausible." They turn to the available types of entities and activities to provide building blocks from which to construct hypothetical mechanisms. "If one knows what kind of activity is needed to do something, then one seeks kinds of entities that can do it, and vice versa" (Machamber, Darden, and Craver 2000, 17).

Mechanisms, therefore, provide a way to solve the pairing problem, and they leave a multitude of traces that can be uncovered if a hypothesized causal relation really exists. For example, those who want to subject Max Weber's hypothesis about the Reformation leading to capitalism do not have to rest content with simply correlating Protestantism with capitalism. They can also look at the detailed mechanism he described for how this came about, and they can look for the traces left by this mechanism (Hedström and Swedberg 1998, 5; Sprinzak 1972).[47]

[45] The points in this paragraph, and the thunderstorm example, come from Dessler (1991).

[46] Jon Elster says: "Are there lawlike generalizations in the social sciences? If not, are we thrown back on mere description and narrative? In my opinion, the answer to both questions is No. The main task of this essay is to explain and illustrate the idea of a *mechanism* as intermediate between laws and descriptions" (Elster 1998, 45).

[47] Hedström and Swedberg (1998) and Sorenson (1998) rightfully criticize causal modeling for ignoring mechanisms and treating correlations among variables as theoretical relationships. But it might be worth remarking that causal modelers in political science have been calling for more theoretical thinking (Achen 1983; Bartels and Brady 1993) for at least two decades, and a constant refrain at the annual meetings of the Political Methodology Group has been the need for better "microfoundations."

7.3 Multiple Causes and Mechanisms

Earlier in this chapter, the need to rule out common causes and to determine the direction of causation in the counterfactual approach led us towards a consideration of multiple causes. In this section, the need to solve the problem of pre-emption and the pairing problem led to a consideration of mechanisms. Together, these approaches lead us to consider multiple causes and the mechanisms that tie these causes together. Many different authors have come to a similar conclusion about the need to identify mechanisms (Cox 1992; Simon and Iwasaki 1988; Freedman 1991; Goldthorpe 2001), and this approach seems commonplace in epidemiology (Hill 1965) where debates over smoking and lung cancer or sexual behavior and AIDS have been resolved by the identification of biological mechanisms that link the behaviors with the diseases.

8 Four Approaches to Causality

8.1 What is Causation?

We are now at the end of our review of four causal approaches. We have described two fundamental features of causality. One is the symmetric association between causes and effects. The other is the asymmetric fact that causes produce effects, but not the reverse. Table 10.1 summarizes how each approach identifies these two aspects of causality.

Regularity and counterfactual approaches do better at capturing the symmetric aspect of causation than its asymmetric aspect. The regularity approach relies upon the constant conjunction of events and temporal precedence to identify causes and effects. Its primary tool is essentially the "Method of Concomitant Variation" proposed by John Stuart Mill in which the causes of a phenomenon are sought in other phenomena which vary in a similar manner. The counterfactual approach relies upon elaborations of the "Method of Difference" to find causes by comparing instances where the phenomenon occurs and instances where it does not occur to see in what circumstances the situations differ. The counterfactual approach suggests searching for surrogates for the closest possible worlds where the putative cause does not occur to see how they differ from the situation where the cause did occur. This strategy leads naturally to experimental methods where the likelihood of the independence of assignment and outcome, which ensures one kind of closeness, can be increased by rigid control of conditions or by randomly assigning treatments to cases. None of these methods is foolproof because none solves the pairing problem or gets at the connections between events, but experimental methods typically offer the best chance of achieving closest possible worlds for comparisons.

Causal approaches that emphasize mechanisms and capacities provide guidance on how to solve the pairing problem and how to get at the connections between

events. Brady and Collier's emphasis upon causal process observations is in that spirit (2004, ch. 13; see also Freedman, this volume). These observations can be thought of as elucidations and tests of possible mechanisms. And the growing interest in mechanisms in the social sciences (Hedström and Swedberg 1998; Elster 1998) is providing a basis for opening up the black box of the Humean regularity and the counterfactual approaches.

The other major feature of causality, the asymmetry of causes and effects, is captured by temporal priority, manipulated events, and the independence of causes. Each notion takes a somewhat different approach to distinguishing causes from effects once the unconditional association of two events (or sets of events) has been established. Temporal priority simply identifies causes with the events that came first. If growth in the money supply reliably precedes economic growth, then the growth in the money supply is responsible for growth. The manipulation approach identifies the manipulated event as the causally prior one. If a social experiment manipulates work requirements and finds that greater stringency is associated with faster transitions off welfare, then the work requirements are presumed to cause these transitions. Finally, one event is considered the cause of another if a third event can be found that satisfies the INUS conditions for a cause and that varies independently of the putative cause. If short circuits vary independently of wooden frame buildings, and both satisfy INUS conditions for burned-down buildings, then both must be causes of those conflagrations. Or if education levels of voters vary independently of their getting the butterfly ballot, and both satisfy INUS conditions for mistakenly voting for Buchanan instead of Gore, then both must be causes of those mistaken votes.

8.2 Causal Inference with Experimental and Observational Data

Now that we know what causation is, what lessons can we draw for doing empirical research? Table 10.1 shows that each approach provides sustenance for different types of studies and different kinds of questions. Table 10.3 presents a "checklist" based on all of the approaches. Regularity and mechanism approaches tend to ask about the causes of effects while counterfactual and manipulation approaches ask about the effects of imagined or manipulated causes. The counterfactual and manipulation approaches converge on experiments, although counterfactual thought experiments flow naturally from the "possible worlds" perspective of the counterfactual appoach. The regularity approach is at home with observational data, and the mechanism approach thrives on analytical models and case studies.

Which method, however, is the best method? Clearly the gold standard for establishing causality is experimental research, but even that is not without flaws. When they are feasible, well-done experiments can help us construct closest possible worlds and explore counterfactual conditions. But we still have to assume that there is no pre-emption occurring which would make it impossible for us to determine the true

Table 10.3. Causality checklist

General Issues
What is the "cause" (C) event? What is the "effect" (E) event?
What is the exact causal statement of how C causes E?
What is the corresponding counterfactual statement about what happens when C does not occur?
What is the causal field? What is the context or universe of cases in which the cause operates?
Is this a physical or social phenomenon or some mixture?
What role, if any, does human agency play?
What role, if any, does social structure play?
Is the relationship deterministic or probabilistic?
Neo-Humean Approach
Is there a constant conjunction (i.e. correlation) of cause and effect?
Is the cause necessary, sufficient, or INUS?
What are other possible causes, i.e. rival explanations?
Is there a constant conjunction after controls for these other causes are introduced?
Does the cause precede the effect? In what sense?
Counterfactual Approach
Is this a singular conjunction of cause and effect?
Can you describe a closest possible (most similar) world to where C causes E but C does not occur? How close are these worlds?
Can you actually observe any cases of this world (or something close to it, at least on average)? Again, how close are these worlds?
In this closest possible world, does E occur in the absence of C?
Are there cases where E occurs but C does not occur? What factor intervenes and what does this tell us about C causing E?
Manipulation Approach
What does it mean to manipulate your cause? Be explicit. How would you describe the cause?
Do you have any cases where C was actually manipulated? How? What was the effect?
Is this manipulation independent of other factors that influence E?
Mechanism and Capacities Approaches
Can you explain, at a lower level, the mechanism(s) by which C causes E?
Do the mechanisms make sense to you?
What other predictions does this mechanism lead to?
Does the mechanism solve the pairing problem?
Can you identify some capacity that explains the way the cause leads to the effect?
Can you observe this capacity when it is present, and measure it?
What other outcomes might be predicted by this capacity?
What are possible pre-empting causes?

impact of the putative cause, and we also have to assume that there are no interactions across units in the treatment and control groups and that treatments can be confined to the treated cases. If, for example, we are studying the impact of a skill training program on the tendency for welfare recipients to get jobs, we should be aware that a very strong economy might pre-empt the program itself and cause those in both the

control and treatment conditions to get jobs simply because employers did not care much about skills. As a result, we might conclude that skills do not count for much in getting jobs even though they might matter a lot in a less robust economy. Or if we are studying electoral systems in a set of countries with a strong bimodal distribution of voters, we should know that the voter distribution might pre-empt any impact of the electoral system by fostering two strong parties. Consequently, we might conclude that single-member plurality systems and proportional representation systems both led to two parties, even though this is not generally true. And if we are studying some educational innovation that is widely known, we should know that teachers in the "control" classes might pick up and use this innovation thereby nullifying any effect it might have.

If we add an investigation of mechanisms to our experiments, we might be able to develop safeguards against these problems. For the welfare recipients, we could find out more about their job search efforts, for the party systems we could find out about their relationship to the distribution of voters, and for the teachers we could find out about their adoption of new teaching methods.

Once we go to observational studies, matters get much more complicated. Spurious correlation is a real danger. There is no way to know whether those cases which get the treatment and those which do not differ from one another in other ways. It is very hard to be confident that the requirements for an experiment hold which are outlined in the next section (and in Campbell and Stonley 1966 and Cook and Campbell 1979). Because nothing has been manipulated, there is no surefire way to determine the direction of causation. Temporal precedence provides some information about causal direction, but it is often hard to obtain and interpret it.

9 Going beyond the Neyman–Rubin–Holland Conditions for Causation

9.1 The Neyman–Rubin–Holland (NRH) Theory

Among statisticians, the best-known theory of causality developed out of the experimental tradition. The roots of this perspective are in Fisher (1926) and especially Neyman ([1923] 1990), and it has been most fully articulated by Rubin (1974; 1978) and Holland (1986). In this section, which is more technical than the rest of this chapter, we explain this perspective, and we evaluate it in terms of the four approaches to causality.

There are four aspects of the Neyman–Rubin–Holland (NRH) approach which can be thought of as developing a recipe for solving the causal inference problem by comparing similar possible worlds, if certain assumptions hold. This approach

consists of a definition, two assumptions, and a method for satisfying one of the two assumptions:

1. A Counterfactual Definition of Causal Effect—Causal relationships are defined using a counterfactual perspective which focuses on estimating causal effects. This definition alone provides no guidance on how researchers can actually identify causes because it relies upon an unobservable counterfactual. To the extent that the NRH approach considers causal priority, it equates it with temporal priority.

2. An Assumption for Creating Comparable Mini-possible Worlds—Non-interference of Units (SUTVA)—Even if we could observe the outcome for some unit (a person or a country) of both the world with the cause present and without the cause, it is possible that the causal effect would depend upon whether other units received the treatment or did not receive the treatment. For example, the impact of a training program on a child in a family might be different when the child and her sibling received the treatment than when the child alone received the treatment. If this kind of thing happens, then it is very hard to define uniquely what we mean by a "causal effect" because there might be some "interference" across units depending upon which units got the treatment and which did not. The NRH counterfactual possible worlds approach assumes that this kind of interference does not occur by making the Stable Unit Treatment Value Assumption (SUTVA) that treats cases as separate, isolated, closest possible worlds which do not interfere or communicate with one another.

3. An Assumption that Finds a Substitute for Insuring the Identicality of the Counterfactual Situation: The Independence of Assignment and Outcome—The counterfactual possible worlds approach not only assumes that units do not interfere with one another, it also assumes that a world identical to our own, except for the existence of the putative cause, can be imagined. The NHR approach goes on to formulate a set of epistemological assumptions, namely the independence of the assignment of treatment and the outcome or the mean conditional independence of assignment and outcome, that make it possible to be sure that two sets of cases, treatments and controls, only differ on average in whether or not they got the treatment.

4. Methods for Insuring Independence of Assignment and Outcome if SUTVA holds—Finally, the NRH approach describes methods such as unit homogeneity or random assignment for obtaining independence or mean independence of assignment and outcome as long as SUTVA holds.

The definition of a causal effect based upon unobserved counterfactuals was first described in a 1923 paper published in Polish by Jerzy Neyman (1990). Although Neyman's paper was relatively unknown until 1990, similar ideas informed much of the statistical work on experimentation from the 1920s to the present. Rubin (1974; 1978; 1990) and Heckman (1979) were the first to stress the importance of independence of assignment and outcome. A number of experimentalists identified the need for the SUTVA assumption (e.g. Cox 1958). Random assignment as a method for estimating

causal effects was first championed by R. A. Fisher in 1925 and 1926. Holland (1986) provides the best synthesis of the entire perspective.

The counterfactual definition of causality rests on the notion of comparing a world with the treatment to a world without it. The fundamental problem of counterfactual definitions of causation is the tension between finding a suitable definition of causation that controls for confounding effects and finding a suitable way of detecting causation given the impossibility of getting perfect counterfactual worlds. As we shall show, the problem is one of relating a *theoretical* definition of causality to an *empirical* one.

9.2 Ontological Definition of Causal Effect Based upon Counterfactuals

Consider a situation in which there is one "unit" A which can be manipulated in some way. Table 10.4 summarizes the situation. Assume that there are two possible manipulations Z_A of the unit, the "control" which we denote by $Z_A = 0$ and the "treatment" which we denote by $Z_A = 1$. Outcomes Y_A are a function $Y_A(Z_A)$ of these manipulations so that the outcome for the control manipulation is $Y_A(0)$ and the outcome for the treatment manipulation is $Y_A(1)$.

According to the NRH understanding of causation, establishing a causal relationship between a treatment Z_A and an outcome $Y_A(Z_A)$ consists of comparing outcomes for the case where the case gets the treatment $Z_A = 1$ and where it does not $Z_A = 0$. Thus we compare:

(a) the value of the outcome variable Y_A for a case that has been exposed to a treatment $Y_A(1)$ with
(b) the value of the outcome variable *for the same case if that case had not been exposed to the treatment* $Y_A(0)$.

In this case, we can define causal impact as follows:

$$E_A = \text{Causal Effect on } A = Y_A(1) - Y_A(0). \tag{1}$$

Note that (a) refers to an actual observation in the treatment condition ("a case that has been exposed to a treatment") so the value $Y_A(1)$ is observed while (b) refers to a counterfactual observation of the control condition ("if that case had not been exposed to the treatment").[48] Because the case was exposed to the treatment, it cannot simultaneously be in the control condition, and the value $Y_A(0)$ is the outcome in the closest possible world where the case was not exposed to the treatment. Although this

[48] For simplicity, we assume that the treatment case has been observed, but the important point is not that the treatment is observed but rather that only one of the two conditions can be observed. There is no reason why the situation could not be reversed with the actual observation of the case in the control group and the counterfactual involving the unobserved impact of the treatment condition.

Table 10.4. Possible worlds, outcomes, and causal effects from manipulation Z for one unit A

	Z_A—Manipulation for Unit A	
POSSIBLE WORLDS:	0 Control	Treatment 1
	$Y_A(0)$	$Y_A(1)$
	Outcomes: $Y_A(Z_A)$	
	Causal Effect: $Y_A(1) - Y_A(0)$	
	Problem: Only one world observable.	

value cannot be observed, we can still describe the conclusions we would draw if we could observe it.

The causal effect E_A for a particular case is the difference in outcomes, $E_A = Y_A(1) - Y_A(0)$, for the case, and if this difference is zero (i.e. if $E_A = 0$), we say the treatment has no net effect.[49] If this difference is nonzero (i.e. E_A is not 0), then the treatment has a net effect. Then, based on the counterfactual approach of David Lewis, there is a causal connection between the treatment and the outcome if two conditions hold. First, the treatment must be associated with a net effect, and second the absence of the treatment must be associated with no net effect.[50]

Although the satisfaction of these two conditions is enough to demonstrate a causal connection, it is not enough to determine the direction of causation or to rule out a common cause. If the two conditions for a causal connection hold, then the third Lewis condition, which establishes the direction of causation and which rules out common cause, cannot be verified or rejected with the available information. The third Lewis condition requires determining whether or not the cause occurs in the closest possible world in which the net effect does not occur. But the only observed world in which the net effect does not occur in the NRH setup is the control

[49] Technically, we mean that the treatment has no effect with respect to that outcome variable.

[50] With a suitable definition of effect, one of these conditions will always hold by definition and the other will be determinative of the causal connection. The NRH approach focuses on the Effect of the Treatment ($E = Y(1) - Y(0)$) in which the control outcome $Y(0)$ is the baseline against which the treatment outcome $Y(1)$ is compared. A nonzero E implies the truth of the counterfactual "if the treatment occurs, then the net effect occurs," and a zero E implies that the counterfactual is false. In the NRH setup the Effect for the Control (EC) must always be zero because $EC = (Y(0) - Y(0))$ is always zero. Hence, the counterfactual "if the treatment is absent then there is no net effect" is always true. The focus on the effect of the treatment (E) merely formalizes the fact that in *any* situation one of the two counterfactuals required for a causal connection can always be defined to be true by an appropriate definition of an effect. Philosophers, by custom, tend to focus on the situation where some effect is associated with some putative cause so that it is always true that "if the cause occurs then the effect occurs as well" and the important question is the truth or falsity of "if the cause does not occur then the effect does not occur." Statisticians such as NRH, with their emphasis on the null hypothesis, seem to prefer the equivalent, but reverse, setup where the important question is the truth or falsity of "if the treatment occurs, then the effect occurs." The bottom line is that a suitable definition of effect can always lead to the truth of one of the two counterfactuals so that causal impacts must always be considered comparatively.

condition in which the cause does not occur by *design* so that there is no way to determine whether suppressing the effect would or would not suppress the cause. There is no way to test the third Lewis condition and to show that the treatment causes the net effect.

Alternatively, the direction of causation can be determined (although common cause cannot be ruled out) if the treatment is manipulated to produce the effect. Rubin and his collaborators mention manipulation when they say that "each of the T treatments must consist of a series of actions that could be applied to each experimental unit" (Rubin 1978, 39) and "it is critical that each unit be *potentially exposable* to any one of the causes" (Holland 1986, 946), but their use of phrases such as "could be applied" or "potentially exposable" suggests that they are more concerned about limiting the possible types of causes than with distinguishing causes from effects.[51] To the degree that causal priority is mentioned in the NRH literature, it is established by temporal precedence. Rubin (1974, 689), for example, says that the causal effect of one treatment over another "for a particular unit and an interval t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to [one treatment] initiated at time t_1 and what would have happened at t_2 if the unit had been exposed to [another treatment] at t_1." Holland (1986, 980) says that "The issue of temporal succession is shamelessly embraced by the model as one of the defining characteristics of a response variable. The idea that an effect might precede a cause in time is regarded as meaningless in the model, and apparently also by Hume." The problem with this approach, of course, is that it does not necessarily rule out common cause and spurious correlation.[52] In fact one of the limitations and possible confusions produced by the NRH approach is its failure to deal with the need for more information to rule out common causes and to determine the direction of causality.

9.3 Finding a Substitute for the Counterfactual Situation: The Independence of Assignment and Outcome

As with the Lewis counterfactual approach, the difficulty with the NRH definition of causal connections is that there is no way to observe both $Y_A(1)$ and $Y_A(0)$ for any particular case. The typical response to this problem is to find two units A and B which are as similar as possible and to consider various possible allocations of the control and the treatment to the two units. (We shall say more about how to ensure this similarity later; for the moment, simply assume that it can be accomplished.) The

[51] Rubin and Holland believe in "NO CAUSATION WITHOUT MANIPULATION" (Holland 1986, 959), which seems to eliminate attributes such as sex or race as possible causes, although Rubin softens this perspective somewhat by describing ways in which sex might be a manipulation (Rubin 1986, 962). Clearly, researchers must consider carefully in what sense some factors can be considered causes.

[52] Consider, for example, an experiment in which randomly assigned special tutoring first causes a rise in self-esteem and then an increase in test scores, but the increase in self-esteem does not cause the increase in test scores. The NRH framework would incorrectly treat self-esteem as the cause of the increased test scores because self esteem is randomly assigned and it precedes and is associated with the rise in test scores. Clearly something more than temporal priority is needed for causal priority.

Table 10.5. Possible worlds, outcomes, and causal effects from manipulations Z for two units A and B

Manipulations for each unit	FOUR POSSIBLE WORLDS			
	$Z_A = 0$, *Control*		$Z_A = 1$, *Treatment*	
	$Z_B = 0$, *Control*	$Z_B = 1$, *Treatment*	$Z_B = 0$, *Control*	$Z_B = 1$, *Treatment*
Outcome value $Y_i(Z_A, Z_B)$, for $i = A$ or B	$Y_A(0,0)$ $Y_B(0,0)$	$Y_A(0,1)$ $Y_B(0,1)$	$Y_A(1,0)$ $Y_B(1,0)$	$Y_A(1,1)$ $Y_B(1,1)$

goal is ultimately to define causal impact as the difference between what happens to A and to B when one of them gets the treatment and the other does not. But, as we shall see, this leads to fundamental problems regarding the definition of causality.

The manipulation for unit A is described by $Z_A = 0$ or $Z_A = 1$ and the manipulation for unit B is described by $Z_B = 0$ or $Z_B = 1$. Table 10.5 illustrates the four possible worlds that could occur based upon the four ways that the manipulations could be allocated. In the first column, both A and B are given the control. In the second column, A gets the control and B gets the treatment. In the third column, A gets the treatment and B gets the control, and in the fourth column, both units get the treatment. The outcomes for these combinations of manipulations are described by $Y_A(Z_A, Z_B)$ and $Y_B(Z_A, Z_B)$.

For each unit, there are then four possible outcome quantities. For example, for A there are $Y_A(0, 0)$, $Y_A(0, 1)$, $Y_A(1, 0)$, and $Y_A(1, 1)$. Similarly for B there are $Y_B(0, 0)$, $Y_B(0, 1)$, $Y_B(1, 0)$, and $Y_B(1, 1)$. For each unit, there are six possible ways to take these four possible outcome quantities two at a time to define a difference that could be considered the causal impact of Z_A, but not all of them make sense as a definition of the causal impact of Z_A. The six possibilities are listed in Table 10.6.

Table 10.6. Six possible definitions of causal impact on unit A

Four observable quantities: $Y_A(0,0)$, $Y_A(0,1)$, $Y_A(1,0)$, $Y_A(1,1)$	
Possible definitions:	
$Y_A(0,0) - Y_A(0,1)$ $Y_A(1,0) - Y_A(1,1)$	Problem: No manipulation of A.
$Y_A(1,1) - Y_A(0,0)$ $Y_A(1,0) - Y_A(0,1)$	Problem: Different treatments for B.
$Y_A(1,0) - Y_A(0,0) = E_A(Z_B = 0)$ $Y_A(1,1) - Y_A(0,1) = E_A(Z_B = 1)$	Both good.

For example, each of $[Y_A(0, 0) - Y_A(0, 1)]$ and $[Y_A(1, 0) - Y_A(1, 1)]$ involves a difference where Z_A does not even vary—in the first case Z_A is the control manipulation for both states of the world and in the second case Z_A is the treatment manipulation. Neither of these differences makes much sense as a definition of the causal impact of Z_A.

Two other pairs of differences, $[Y_A(1, 1) - Y_A(0, 0)]$ and $[Y_A(1, 0) - Y_A(0, 1)]$, seem better insofar as they each involve differences in which A received the treatment in one case and the control in the other case, but the manipulation of B differs within each pair. In the first difference, for example, we are comparing the outcome for A in the world in which A gets the treatment and B does not with the world in which A does not get the treatment and B gets it. At first blush, it might seem that it doesn't really matter what happens to B, but a moment's reflection suggests that unless A and B *do not interfere* with one another, it might matter a great deal what happens to B.

Suppose, for example, that A and B are siblings, adjacent plots of land, two students in the same class, two people getting a welfare program in the same neighborhood, two nearby countries, or even two countries united by common language and traditions. Then for treatments as diverse as new teaching methods, propaganda, farming techniques, new scientific or medical procedures, new ideas, or new forms of government it might matter for the A member of the pair what happens to the B member because of causal links between them. For example, if a sibling B is given a special educational program designed to increase achievement, it seems possible that some of this impact will be communicated to the other sibling A, even when A does not get the treatment directly. Or if a new religion or religious doctrine is introduced into one country, it seems possible that it will have an impact on the other country. In both cases, it seems foolish to try to compare the impact of different manipulations of A when different things have also been done to B, unless we can be sure that a manipulation of B has no impact on A or unless we define the manipulation of B as part of the manipulation of A.

This second possibility deserves some comment. If the manipulation of B is part of the manipulation of A, then we really have not introduced a new unit when we decided to consider B as well as A. In this situation we can think of the differences listed above, $[Y_A(1, 1) - Y_A(0, 0)]$ and $[Y_A(1, 0) - Y_A(0, 1)]$, as indicating the impact on A of the manipulation of the combined unit $A + B$. For the first difference, $[Y_A(1, 1) - Y_A(0, 0)]$, the manipulation consists of applying $Z_A = 1$ and $Z_B = 1$ as the treatment to $A + B$ and the $Z_A = 0$ and $Z_B = 0$ as the control to $A + B$. Similar reasoning applies to the second difference, $[Y_A(1, 0) - Y_A(0, 1)]$. There are two lessons to be learned from this discussion. First, it is not as easy as it might seem to define isolated units, and the definition of separate units partly depends upon how they will be affected by the manipulation. Second, it does not make much sense to use $[Y_A(1, 1) - Y_A(0, 0)]$ or $[Y_A(1, 0) - Y_A(0, 1)]$ as the definition of the causal impact of the treatment Z_A on A.

This leaves us with the following pairs which are plausible definitions of the causal effect for each unit, depending upon what happens to the other unit. These pairs are

Table 10.7. Theoretical definitions summarized for units A and B

For unit A:
$Y_A(1,0) - Y_A(0,0) = E_A(Z_B = 0)$
$Y_A(1,1) - Y_A(0,1) = E_A(Z_B = 1)$
For unit B:
$Y_B(0,1) - Y_B(0,0) = E_B(Z_A = 0)$
$Y_B(1,1) - Y_B(1,0) = E_B(Z_A = 1)$

summarized in Table 10.7. For example, for A:

$$E_A(Z_B = 0) = Y_A(1, 0) - Y_A(0, 0), \quad \text{and} \tag{2}$$
$$E_A(Z_B = 1) = Y_A(1, 1) - Y_A(0, 1).$$

And for B we have:

$$E_B(Z_A = 0) = Y_B(0, 1) - Y_B(0, 0), \quad \text{and} \tag{3}$$
$$E_B(Z_A = 1) = Y_B(1, 1) - Y_B(1, 0).$$

Consider the definitions for A in (2). Both definitions seem sensible because each one takes the difference between the outcome when A is treated and the outcome when A is not treated, but they differ on what happens to B. In the first case, B is given the control manipulation and in the second case, B is given the treatment manipulation. From the preceding discussion, it should be clear that these might lead to different sizes of effects. The impact of a pesticide on a plot A, for example, might vary dramatically depending upon whether or not the adjacent plot B got the pesticide. The effect of a propaganda campaign might vary dramatically depending upon whether or not a sibling got the propaganda message. As a result, there is no a priori reason why $E_A(Z_B = 0)$ and $E_A(Z_B = 1)$ should be the same thing. The impact on A of a treatment might depend upon what happens to B.

One response to this problem might be simply to agree that $E_A(Z_B = 0)$ and $E_A(Z_B = 1)$ (and $E_B(Z_A = 0)$ and $E_B(Z_A = 1)$) are different and that a careful researcher would want to measure both of them. But how could that be done? Neither can be measured directly because each requires that the unit A both get and not get the treatment, which is clearly impossible. In terms of our notation, the problem is that each difference above involves different values for Z_A and Z_B. For example, $E_A(Z_B = 0)$ which equals $Y_A(1, 0) - Y_A(0, 0)$ involves one state of the world where A gets the treatment and B does not and another state of the world where A does not get the treatment and B does not. Both states of the world cannot occur.

Table 10.8. Observationally feasible definitions of causality

Four states of the world and four possible definitions:
(1) $\{Z_A = 1$ and $Z_B = 1\}$
Observe $Y_A(1,1)$ and $Y_B(1,1) \rightarrow$ Difference Zero
(2) $\{Z_A = 0$ and $Z_B = 0\}$
Observe $Y_A(0,0)$ and $Y_B(0,0) \rightarrow$ Difference Zero
(3) $\{Z_A = 1$ and $Z_B = 0\}$
Observe $Y_A(1,0)$ and $Y_B(1,0) \rightarrow E^*(1,0) = Y_A(1,0) - Y_B(1,0)$
(4) $\{Z_A = 0$ and $Z_B = 1\}$
Observe $Y_A(0,1)$ and $Y_B(0,1) \rightarrow E^*(0,1) = Y_A(0,1) - Y_B(0,1)$

9.4 Observable Definitions of Causality

As noted earlier, the standard response to this problem is to consider definitions of causal impact that are observable because the relevant quantities can be measured in the same state of the world—thus avoiding the problem of making comparisons across multiple worlds or between the existing world and another, "impossible," world. With two units and a dichotomous treatment, four states of the world are possible: $\{Z_A = 1$ and $Z_B = 1\}$, $\{Z_A = 0$ and $Z_B = 0\}$, $\{Z_A = 1$ and $Z_B = 0\}$, and $\{Z_A = 0$ and $Z_B = 1\}$. These are listed in Table 10.5 along with the two observable quantities, Y_A and Y_B, one for A and one for B, for each state of the world.

The four differences of these two quantities are listed in Table 10.8. Each difference is a candidate to be considered as a measure of causal impact. The differences for the first and second of these four states of the world do not offer much opportunity for detecting the causal impact of Z because there is no variability in the treatment between the two units.[53] Consequently, we consider the differences for the third and fourth cases.

For the state of the world $\{Z_A = 1$ and $Z_B = 0\}$ we can compute the following based upon observable quantities:

$$E^*(1,0) = Y_A(1,0) - Y_B(1,0), \tag{4}$$

where the difference involves terms that occur together in one state of the world. Note that we denote this empirical definition of causality by an asterisk. This difference is computable, but does it represent a causal impact? Intuitively, the problem with using it as an estimate of causal impact is that A and B might be quite different to begin with. Suppose we are trying to estimate the impact of a new teaching method. Person A might be an underachiever while person B might be an overachiever. Hence, even if the method works, person A might score lower on a test after treatment than person B, and the method will be deemed a failure. Or suppose we are trying to determine

[53] Consider, for example, the difference $E^*(1,1) = Y_A(1,1) - Y_B(1,1)$ for state of the world $\{Z_A = 1$ and $Z_B = 1\}$. If we make the very reasonable assumption of identicality described below, then $Y_B(1,1) = Y_A(1,1)$ so that $E^*(1,1)$ is always zero which is not a very interesting "causal effect." The same result applies to the state of the world $\{Z_A = 0$ and $Z_B = 0\}$.

the impact of a new voting machine. County A might be very competent at running elections while county B might not be. Consequently, even if the machine works badly, county A with the new system might perform better than county B without it—once again leading to the wrong inference. Clearly $E^*(1, 0)$ alone is not a very good definition of causal impact. One of the problems is that preexisting differences between the units can confound causal inference.

How, then, can $E^*(1, 0)$ be used to make a better causal inference? Surveying the four definitions of causal impact in equations (2) and (3) above, this definition seems most closely related to two of them:

$$E_A(Z_B = 0) = Y_A(1, 0) - Y_A(0, 0), \quad \text{and} \tag{5a}$$

$$E_B(Z_A = 1) = Y_B(1, 1) - Y_B(1, 0). \tag{5b}$$

Consider the first of these, $E_A(Z_B = 0)$. Obviously, $E^*(1, 0)$ will equal $E_A(Z_B = 0)$ if the second term in the expression for $E^*(1, 0)$ which is $Y_B(1, 0)$ equals the second term in the expression for $E_A(Z_B = 0)$ which is $Y_A(0, 0)$. Thus we require that:

$$Y_B(1, 0) = Y_A(0, 0). \tag{6}$$

What conditions will ensure that this is so?

We shall make the transformation of $Y_B(1, 0)$ into $Y_A(0, 0)$ in two steps which are depicted on Table 10.9. If A and B are identical and Z_A and Z_B are identical as well[54] (although we haven't indicated how this might be brought about yet) it might be reasonable to suppose that:

$$Y_B(1, 0) = Y_A(0, 1) [\textit{Identicality of units and treatment or Unit Homogeneity}]. \tag{7}$$

That is, A and B are mirror images of one another so that the impact of $Z_A = 1$ and $Z_B = 0$ on B is the same as the impact of $Z_A = 0$ and $Z_B = 1$ on A.

This assumption is the same as what Holland (1986) calls "unit homogeneity" in which units are prepared carefully "so that they 'look' identical in all relevant aspects" (Holland 1986, 948). This assumption is commonly made in laboratory work where identical specimens are tested or where the impacts of different manipulations are studied for the identical setup. It obviously requires a great deal of knowledge about what makes things identical to one another and an ability to control these factors. It is typically not a very good assumption in the social sciences.

With this assumption, $E^*(1, 0) = Y_A(1, 0) - Y_A(0, 1)$ which is a definition of causality that we discarded earlier because of the possibility that if B gets the treatment when A does not, then A will be affected even when A does not get the treatment. We discarded this definition because, for example, the impact $Y_A(0, 1)$ of the treatment on Amy when Beatrice gets the treatment might be substantial—perhaps

[54] By saying that Z_A and Z_B have to be comparable, we mean that $Z_A = 0$ and $Z_B = 0$ are the same thing and $Z_A = 1$ and $Z_B = 1$ are the same thing.

Table 10.9. Linking observational data to theoretical definitions of causality through unit identicality and noninterference of units

Observational	Unit identicality [unit homogeneity]	Noninterference of units [SUTVA]	Theoretical definition
$E^{-}(1,0) = Y_A(1,0) - Y_B(1,0)$	$Y_B(1,0) = Y_A(0,1) \rightarrow Y_A(1,0) - Y_A(0,1)$	$Y_A(0,1) = Y_A(0,0) \rightarrow Y_A(1,0) - Y_A(0,0)$ $Y_A(1,0) = Y_A(1,1) \rightarrow Y_A(1,1) - Y_A(0,1)$	$E_A(Z_B = 0)$ $E_A(Z_B = 1)$
	$Y_A(1,0) = Y_B(0,1) \rightarrow Y_B(0,1) - Y_B(1,0)$	$Y_B(1,0) = Y_B(0,0) \rightarrow Y_B(0,1) - Y_B(0,0)$ $Y_B(0,1) = Y_B(1,1) \rightarrow Y_B(1,1) - Y_B(1,0)$	$E_B(Z_A = 0)$ $E_B(Z_A = 1)$
$E^{+}(0,1) = Y_B(0,1) - Y_A(0,1)$	$Y_B(0,1) = Y_A(1,0) \rightarrow Y_A(1,0) - Y_A(0,1)$	$Y_A(0,1) = Y_A(0,0) \rightarrow Y_A(1,0) - Y_A(0,0)$ $Y_A(1,0) = Y_A(1,1) \rightarrow Y_A(1,1) - Y_A(0,1)$	$E_A(Z_B = 0)$ $E_A(Z_B = 1)$
	$Y_A(0,1) = Y_B(1,0) \rightarrow Y_B(0,1) - Y_B(1,0)$	$Y_B(1,0) = Y_B(0,0) \rightarrow Y_B(0,1) - Y_B(0,0)$ $Y_B(0,1) = Y_B(1,1) \rightarrow Y_B(1,1) - Y_B(1,0)$	$E_B(Z_A = 0)$ $E_B(Z_A = 1)$

as much as when Amy gets the treatment alone which is $Y_A(1, 0)$. In that case, $E^*(1, 0)$ seems like a poor definition of the causal impact of Z_A when what we really want is the definition in (5a) above. But to get to that definition, we must suppose that:

$$Y_A(0, 1) = Y_A(0, 0) \text{ [Non-interference of units or SUTVA]}. \tag{8}$$

In effect, this requires that we believe that the causal impact of manipulation Z_A on A is not affected by whether or not B gets the treatment. Rubin (1990) calls this the "Stable-Unit-Treatment Value Assumption" (SUTVA). As we have already seen, this is a worrisome assumption, and we shall have a great deal to say about it later.

Similarly, $E^*(1, 0)$ will equal the second definition (5b) above, $E_B(Z_A = 1)$, if the first term in the expression for $E^*(1, 0)$ which is $Y_A(1, 0)$ equals the first term in the expression for $E_B(Z_A = 1)$ which is $Y_B(1, 1)$. Once again, if A and B are identical and Z_A and Z_B are identical then we can suppose that:

$$Y_A(1, 0) = Y_B(0, 1) \text{ [Identicality of units and treatment or unit homogeneity]}. \tag{9}$$

In addition we need to assume that the causal impact of manipulation Z_A on B is not affected by whether or not A gets the treatment:

$$Y_B(0, 1) = Y_B(1, 1) \text{ [Noninterference of Units or SUTVA]}. \tag{10}$$

To summarize, to get a workable operational definition of causality, we need to assume that one of the following holds true:

$$Y_B(1, 0) = Y_A(0, 1) = Y_A(0, 0), \text{ or} \tag{11a}$$

$$Y_A(1, 0) = Y_B(0, 1) = Y_B(1, 1). \tag{11b}$$

The first equality in each line holds true if we assume identicality and the second holds true if we assume noninterference (SUTVA). Note that if both (11a) and (11b) are true, then the definitions of $E^*(1, 0)$, $E_A(Z_B = 0)$, and $E_B(Z_A = 1)$ all collapse to one another.

Instead of (4) as the operational definition of causal impact, we might consider the following which is the effect for the state of the world $\{Z_A = 0 \text{ and } Z_B = 1\}$:

$$E^*(0, 1) = Y_B(0, 1) - Y_A(0, 1), \tag{12}$$

where the difference involves terms that occur in only one state of the world. Surveying the four theoretical definitions of causal impact in equations (2) and (3) above, this definition seems most closely related to these two:

$$E_A(Z_B = 1) = Y_A(1, 1) - Y_A(0, 1) \tag{13a}$$

$$E_B(Z_A = 0) = Y_B(0, 1) - Y_B(0, 0), \tag{13b}$$

and these two are the remaining two after the ones in (5) are considered. To make these definitions work, we require, analogously to (11) above, that:

$$Y_B(0, 1) = Y_A(1, 0) = Y_A(1, 1), \text{ or} \tag{14a}$$

$$Y_A(0, 1) = Y_B(1, 0) = Y_B(0, 0). \tag{14b}$$

where as before, the first equality in each line comes from identicality and the second comes from assuming noninterference. Once again, with these assumptions, then the definitions of $E^*(0, 1)$, $E_A(Z_B = 1)$, and $E_B(Z_A = 0)$ collapse into the same thing. And if both (11a,b) and (14a,b) hold, then $E^*(1, 0)$ equals $E^*(0, 1)$, and these definitions are all the same. Table 10.9 summarizes this entire argument.

9.5 Getting around Identicality (Unit Homogeneity) through Average Causal Effect

It is clear that the assumptions of noninterference (SUTVA) and identicality are sufficient to define causality unambiguously, but are they necessary? They are very strong assumptions. Can we do without one or the other? Suppose, for example, that we just assume noninterference so that $Y_A(j, k) = Y_A(j, k')$ and $Y_B(j, k) = Y_B(j, k')$ for $j = 1, 2$ and $k \neq k'$. Then we get the comforting result that the two theoretical definitions of causal impact for A (in (2) above) and the two for B (in (3) above) are identical:

$$E_A(Z_B = 0) = Y_A(1, 0) - Y_A(0, 0) = Y_A(1, 1) - Y_A(0, 1) = E_A(Z_B = 1)$$

$$E_B(Z_A = 0) = Y_B(0, 1) - Y_B(0, 0) = Y_B(1, 1) - Y_B(1, 0) = E_B(Z_A = 1).$$

Table 10.10 depicts this argument (moving from the rightmost column in the table to the second to the right column.) Since these equations hold, we denote the common causal effects as simply E_A and E_B:

$$E_A = E_A(Z_B = 0) = E_A(Z_B = 1)$$

$$E_B = E_B(Z_A = 0) = E_B(Z_A = 1).$$

These assumptions alone, however, will not allow us to link these theoretical definitions with the empirical possibilities $E^*(1, 0)$ and $E^*(0, 1)$. We need some additional assumption such as identicality of A and B which would ensure that $E_A = E_B$.

Can we get around identicality? Consider the following maneuver. Although we cannot observe both $E^*(1, 0)$ and $E^*(0, 1)$ at the same time, consider their average which we shall call the Average Causal Effect or ACE:

$$\begin{aligned}
\text{ACE} &= (1/2)[E^*(1, 0) + E^*(0, 1)] \\
&= (1/2)\{[Y_A(1, 0) - Y_B(1, 0)] + [Y_B(0, 1) - Y_A(0, 1)]\} \\
&= (1/2)\{[Y_A(1, 0) - Y_A(0, 1)] + [Y_B(0, 1) - Y_B(1, 0)]\} \\
&= (1/2)\{[Y_A(1, 0) - Y_A(0, 0)] + [Y_B(0, 1) - Y_B(0, 0)]\}
\end{aligned}$$

Table 10.10. Linking observational data to theoretical definitions of causality through noninterference of units and average causal effect

Observational →	→ Noninterference →	→ Average Causal Effect ← ACE = $[E^+(1,0) + E^+(0,1)]/2$	← Noninterference ←	← Theoretical
$E^+(1,0) = Y_A(1,0) - Y_B(1,0)$	$Y_B(1,0) = Y_B(0,0)$ → $Y_A(1,0) - Y_B(0,0)$ $Y_A(1,0) = Y_A(1,1)$ → $Y_A(1,1) - Y_B(1,0)$	Take first and third on left: ACE = $[Y_A(1,0) - Y_B(0,0) + Y_B(0,1) - Y_A(0,0)]/2$ = $[Y_A(1,0) - Y_A(0,0) + Y_B(0,1) - Y_B(0,0)]/2$ = $[E_A + E_B]$ (Using results from panels to right)	$Y_A(1,0) = Y_A(1,1)$ $Y_A(0,0) = Y_A(0,1)$ Hence: $E_A = E_A(Z_B = 0)$ = $E_A(Z_B = 1)$ = $Y_A(1,\cdot) - Y_A(0,\cdot)$	$E_A(Z_B = 0) = Y_A(1,0) - Y_A(0,0)$ $E_A(Z_B = 1) = Y_A(1,1) - Y_A(0,1)$
$E^-(0,1) = Y_B(0,1) - Y_A(0,1)$	$Y_A(0,1) = Y_A(0,0)$ → $Y_B(0,1) - Y_A(0,0)$ $Y_B(0,1) = Y_B(1,1)$ → $Y_B(1,1) - Y_A(0,1)$	Take second and fourth on left: ACE = $[Y_A(1,1) - Y_B(1,0) + Y_B(1,1) - Y_A(0,1)]/2$ = $[Y_A(1,1) - Y_A(0,1) + Y_B(1,1) - Y_A(1,0)]/2$ = $[E_A + E_B]$ (Using results from panels to right)	$Y_B(0,1) = Y_B(1,1)$ $Y_B(0,0) = Y_B(1,0)$ Hence: $E_B = E_B(Z_A = 0)$ = $E_B(Z_A = 1)$ = $Y_B(\cdot,1) - Y_B(\cdot,0)$	$E_B(Z_A = 0) = Y_B(0,1) - Y_B(0,0)$ $E_B(Z_A = 1) = Y_B(1,1) - Y_B(1,0)$

where the second line uses the definitions of $E^*(1, 0)$ and $E^*(0, 1)$, the third line is simply algebra, and the last line comes from noninterference. This argument is depicted in Table 10.10 as we move from the first to the second to the third column. As a result, we can write:

$$\text{ACE} = (1/2)[E_A + E_B].$$

Therefore, the ACE represents the average causal impact of Z_A on A and Z_B on B. If identicality (of A to B and Z_A to Z_B) held, then ACE would simply be the causal impact of Z.

Unfortunately, we cannot observe ACE, and we do not want to assume identicality. We can observe either $E^*(1, 0)$ or $E^*(0, 1)$, but not both. We can, however, do the following. We can randomly choose the state of the world, either $\{Z_A = 1 \text{ and } Z_B = 0\}$ or $\{Z_A = 0 \text{ and } Z_B = 1\}$. *Randomization in this way ensures that the treatment is assigned at random.* Once we have done this, we can take the observed value of either $E^*(1, 0)$ or $E^*(0, 1)$ as an estimate of ACE. The virtue of this estimate is that it is a statistically unbiased estimate of the average impact of Z_A on A and Z_B on B. That is, in repeated trials of this experiment (assuming that repeated trials make sense), the expected value of ACE will be equal to the true causal effect. Randomization ensures that we don't fall into the trap of confounding because, in repeated trials, there is no relationship between the assignment of treatment and units.

But the measure has two defects. First, it may be problematic to consider the average impact of Z_A on A and Z_B on B if they are not similar kinds of things. Once we drop identicality, it is quite possible that A and B could be quite different kinds of entities, say a sick person (A) and a well person (B). Then one would be randomly chosen to get some medicine, and the subsequent health (Y) of each person would be recorded. If the sick person A got the medicine then the causal effect E_A would be the difference between the health $Y_A(1, 0)$ of the sick person (after taking the medicine) and the health of the well person $Y_B(1, 0)$. If the well person B got the medicine, then the causal effect E_B would be the difference between the health $Y_B(0, 1)$ of the well person (after taking the medicine) and the health of the sick person $Y_A(0, 1)$. If the medicine works all the time and makes people well, then E_A will be zero (giving the medicine to the sick person will make him like the well person) and E_B will be positive (giving the medicine to the well person will not change her but not giving it to the sick person will leave him still sick)—hence the average effect will be to say that the medicine works, half the time. In fact, the medicine works all the time—when the person is sick. More generally, and somewhat ridiculously, A could be a person and B could be a tree, a dog, or anything. Thus, we need some assumption like the identicality of the units in order for our estimates of causal effect to make any sense. One possibility is that they are randomly chosen from some well-defined population to whom the treatment might be applied in the future.

The second defect of the measure is that it is only correct in repeated trials. In the medical experiment described above, if the well person is randomly assigned the medicine, then the experiment will conclude that the medicine does not work. The usual response to this problem is to multiply the number of units so that the random

assignment to treatment group and control group creates groups that are, because of the law of large numbers, very similar, on average. This strategy certainly can make it possible to make statistical statements about the likelihood that an observed difference between the treatment and control groups is due to chance or to some underlying true difference. But it relies heavily upon multiplying the number of units, and it seems that multiplying the number of units brings some risks with it.

9.6 Multiplying the Number of Units and the Noninterference (SUTVA) Assumption

We started this section with a very simple problem in what is called singular causation. We asked: How does manipulation $Z = 1$ affect the outcome Y_A for unit A? Equation (1) provided a very simple definition of what we meant by the causal effect. It is simply $E_A = Y_A(1) - Y_A(0)$. This simple definition foundered because we cannot observe both $Y_A(1)$ and $Y_A(0)$. To solve this problem, we multiplied the number of units. Multiplying the number of units makes it possible to obtain an observable estimate of causal effect by either making the noninterference and identicality assumptions or by making the noninterference assumption and using randomization to achieve random assignment. But these assumptions lead us into the difficulties of defining a population of similar things from which the units are chosen and the problem of believing the noninterference assumption. These problems are related because they suggest that ultimately researchers must rely upon some prior knowledge and information in order to be sure that units or cases can be compared. But how much knowledge is needed? Are these assumptions really problematic? Should we, for example, be worried about units affecting one another?

Yes. Suppose people in a treatment condition are punished for poor behavior while those in a control condition are not. Further suppose that those in the control condition who are "near" (i.e. live in the same neighborhood or communicate regularly with one another) those in the treatment condition are not fully aware that they are exempt from punishment or they fear that they might be made subject to it. Wouldn't their behavior change in ways that it would not have changed if there had never been a treatment condition? Doesn't this mean that it would be difficult, if not impossible, to satisfy the noninterference condition?

In the Cal-Learn experiment in California, for example, teenage girls on welfare in the treatment group had their welfare check reduced if they failed to get passing grades in school. Those in the randomly selected control group were not subject to reductions but many thought they were in the treatment group (probably because they knew people who were in the treatment group) and they appear to have worked to get passing grades to avoid cuts in welfare (Mauldon et al. 2000).[55] Their decision

[55] Experimental subjects were told which group they were in, but some apparently did not get the message. They may not have gotten the message because the control group was only a small number of people and almost all teenage welfare mothers in the state were in the treatment group. In these

to get better grades, however, may have led to an underestimate of the impact of Cal-Learn because it reduced the difference between the treatment group and the control group. The problem here is that there is interaction between the units. To rule out these possibilities, Rubin (1990) proposed the "Stable-Unit-Treatment-Value-Assumption (SUTVA)" which, as we have seen, asserts that the outcome for a particular case does not depend upon what happens to the other cases or which of the supposedly identical treatments the unit receives.

Researchers using human subjects have worried about the possibility of interference. Cook and Campbell (1986, 148) mention four fundamental threats to randomized experiments. Compensatory rivalry occurs when control units decide that even though they are not getting the treatment, they can do as well as those getting it. Resentful demoralization occurs when those not getting the treatment become demoralized because they are not getting the treatment. Compensatory equalization occurs when those in charge of control units decide to compensate for the perceived inequities between treatment and control units, and treatment diffusion occurs when those in charge of control units mimic the treatment because of its supposed beneficial effects.

SUTVA implies that each supposedly identical treatment really is identical and that each unit is a separate, isolated possible world that is unaffected by what happens to the other units. SUTVA is the master assumption that makes controlled or randomized experiments a suitable solution to the problem of making causal inferences. SUTVA ensures that treatment and control units really do represent the closest possible worlds to one another except for the difference in treatment. In order to believe that SUTVA holds, we must have a very clear picture of the units, treatments, and outcomes in the situation at hand so that we can convince ourselves that experimental (or observational) comparisons really do involve similar worlds. Rubin (1986, 962) notes, for example, that statements such as "If the females at firm f had been male, their starting salaries would have averaged 20% higher" require much more elaboration of the counterfactual possibilities before they can be tested. What kind of treatment, for example, would be required for females to be males? Are individuals or the firm the basic unit of analysis? Is it possible simply to randomly assign men to the women's jobs to see what would happen to salaries? From what pool would these men be chosen? If men were randomly assigned to some jobs formerly held by women, would there be interactions across units that would violate SUTVA?

Not surprisingly, if the SUTVA assumption fails, then it will be at best hard to generalize the results of an experiment and at worst impossible to even interpret its results. Generalization is hard if, for example, imposing a policy of welfare time-limits on a small group of welfare recipients has a much different impact than imposing it upon every recipient. Perhaps the imposition of limits on the larger group generates a negative attitude toward welfare that encourages job seeking which is not generated

circumstances, an inattentive teenager in the control group could have sensibly supposed that the program applied to everyone. Furthermore, getting better grades seemingly had the desired effect because their welfare check was not cut!

when the limits are only imposed on a few people. Or perhaps the random assignment of a "Jewish" culture to one country (such as Israel) is much different than assigning it to a large number of countries in the same area. In both cases, the pattern of assignment to treatments seems to matter as much as the treatments themselves because of interactions among the units, and the interpretation of these experiments might be impossible because of the complex interactions among units. If SUTVA does not hold, then there are no ways such as randomization to construct closest possible worlds, and the difficulty of determining closest possible worlds must be faced directly.

If SUTVA holds and if there is independence of assignment and outcome through randomization, then the degree of causal connection can be estimated.[56] But there is no direct test that can ensure that SUTVA holds and there are only partial tests of "balance" to ensure that randomization has been done properly. Much of the art in experimentation goes into strategies that will increase the likelihood that they do hold. Cases can be isolated from one another to minimize interference, treatments can be made as uniform as possible, and the characteristics and circumstances of each case can be made as uniform as possible, but nothing can absolutely ensure that SUTVA and the independence of assignment and outcome hold.[57]

9.7 Summary of the NRH Approach

If noninterference across units (SUTVA) holds and if independence of assignment and outcome hold, then mini-closest-possible worlds have been created which can be used to compare the effects in a treatment and control condition. If SUTVA holds, then there are three ways to get the conditional independence conditions to hold:

(a) Controlled experiments in which identicality (unit homogeneity) holds.
(b) Statistical experiments in which random assignment holds.
(c) Observational studies in which corrections are made for covariates that ensure mean conditional independence of assignment and outcome.

The mathematical conditions required for the third method to work follow easily from the Neyman–Holland–Rubin setup, but there is no method for identifying the proper covariates. And outside of experimental studies, there is no way to be sure that conditional independence of assignment and outcome holds. Even if we know about *something* that may confound our results, we may not know about *all* things, and without knowing all of them, we cannot be sure that correcting for some of them

[56] If SUTVA fails and independence of assignment and outcome obtains, then causal effects can also be estimated, but they will differ depending on the pattern of treatments. Furthermore, the failure of SUTVA may make it impossible to rely on standard methods such as experimental control or randomization to ensure that the independence of assignment and outcome holds because the interaction of units may undermine these methods.

[57] Although good randomization can make it very likely that there is independence of assignment and outcome.

ensures conditional independence. Thus observational studies face the problem of identifying a set of variables that will ensure conditional independence so that the impact of the treatment can be determined. A great deal of research, however, does this in a rather cavalier way.

Even if SUTVA and some form of conditional independence is satisfied, the NRH framework, like Lewis's counterfactual theory to which it is a close relative, can only identify causal connections. Additional information is needed to rule out spurious correlation and to establish the direction of causation. Appeal can be made to temporal precedence or to what was manipulated to pin down the direction of causation, but neither of these approaches provides full protection against common cause. More experiments or observations which study the impact of other variables which suppress supposed causes or effects may be needed, and these have to be undertaken imaginatively in ways that explore different possible worlds.

References

Abbott, A. 1983. Sequences of social events. *Historical Methods*, 16: 129–47.

—— 1992. From causes to events. *Sociological Methods and Research*, 20: 428–55.

—— 1995. Sequence analysis: new methods for old ideas. *Annual Review of Sociology*, 21: 93–113.

Achen, C. H. 1983. Toward theories of data: the state of political methodology. In *Political Science: The State of the Discipline*, ed. A. Finifter. Washington, DC: American Political Science Association.

Bartels, L., and Brady, H. E. 1993. The state of quantitative political methodology. In *Political Science: The State of the Discipline*, 2nd edn., ed. A. Finifter. Washington, DC: American Political Science Association.

Beauchamp, T. L., and Rosenberg, A. 1981. *Hume and the Problem of Causation*. New York: Oxford University Press.

Bennett, J. 1988. *Events and Their Names*. Indianapolis: Hackett.

Berger, P. L., and Luckmann, T. 1966. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Garden City, NY: Anchor.

Brady, H. E., and Collier, D. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. New York: Rowman and Littlefield.

—— Herron, M. C., Mebane, W. R., Sekhon, J. S., Shotts, W. S., and Wand, J. 2001. Law and data: the butterfly ballot episode. *PS: Political Science and Politics*, 34: 59–69.

Campbell, D. T., and Stanley, J. C. 1966. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cartwright, N. 1989. *Nature's Capacities and Their Measurement*. New York: Oxford University Press.

Cook, T. D., and Campbell, D. T. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

—— —— 1986. The causal assumptions of quasi-experimental practice. *Synthese*, 68: 141–180.

Cox, D. R. 1958. *The Planning of Experiments*. New York: Wiley.

—— 1992. Causality: some statistical aspects. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 155: 291–301.

Cox, G. W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems*, New York: Cambridge University Press.

Davidson, D. 2001. *Essays on Actions and Events*, 2nd edn. Oxford: Clarendon Press.

Dessler, D. 1991. Beyond correlations: toward a causal theory of war. *International Studies Quarterly*, 35: 337–355.

Dilthey, W. 1961. *Pattern and Meaning in History: Thoughts on History and Society*. New York: Harper.

Durkheim, E. 1982 [1895]. *The Rules of Sociological Method*. New York: Free Press.

Elster, J. 1998. A plea for mechanisms. In *Social Mechanisms*, ed. P. Hedström and R. Swedberg. Cambridge: Cambridge University Press.

Ehring, D. 1997. *Causation and Persistence: A Theory of Causation*. New York: Oxford University Press.

Fearon, J. D. 1991. Counterfactuals and hypothesis testing in political science. *World Politics*, 43: 169–95.

Fisher, R. A., Sir 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

—— 1926. The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33: 503–13.

—— 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

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

—— 1991. Statistical models and shoe leather. *Sociological Methodology*, 21: 291–313.

—— 1997. From association to causation via regression. Pp. 113–61 in *Causality in Crisis?* ed. V. R. McKim and S. P. Turner, Notre Dame, Ind.: University of Notre Dame Press.

—— 1999. From association to causation: some remarks on the history of statistics. *Statistical Science*, 14: 243–58.

Gasking, D. 1955. Causation and recipes. *Mind*, 64: 479–87.

Glennan, S. S. 1996. Mechanisms and the nature of causation. *Erkenntnis*, 44: 49–71.

Goldthorpe, J. H. 2001. Causation, statistics, and sociology. *European Sociological Review*, 17: 1–20.

Goodman, N. 1947. The problem of counterfactual conditionals. *Journal of Philosophy*, 44: 113–28.

Harré, R., and Madden, E. H. *c* 1975. *Causal Powers: A Theory of Natural Necessity*. Oxford: B. Blackwell.

Hausman, D. M. 1998. *Causal Asymmetries*. New York: Cambridge University Press.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica*, 47: 153–62.

Hedström, P., and Swedberg, R. (eds.) 1998. *Social Mechanisms: An Analytical Approach to Social Theory*. New York: Cambridge University Press.

Hempel, C. G. 1965. *Aspects of Scientific Explanation*. New York: Free Press.

Hill, A. B. 1965. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58: 295–300.

Holland, P. W. 1986. Statistics and causal inference (in theory and methods). *Journal of the American Statistical Association*, 81: 945–60.

Hume, D. 1739. *A Treatise of Human Nature*, ed. L. A. Selby-Bigge and P. H. Nidditch. Oxford: Clarendon Press.

—— 1748. *An Enquiry Concerning Human Understanding*, ed. T. L. Beauchamp. New York: Oxford University Press.

Kitcher, P., and Salmon, W. 1987. Van Fraassen on explanation. *Journal of Philosophy*, 84: 315–30.

Lakoff, G. and Johnson, M. 1980*a*. Conceptual metaphor in everyday language. *Journal of Philosophy*, 77 (8): 453–86.

——— 1980*b*. *Metaphors We Live By*. Chicago: University of Chicago Press.

——— 1999. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.

Lewis, D. 1973*a*. *Counterfactuals*. Cambridge, Mass: Harvard University Press.

—— 1973*b*. Causation. *Journal of Philosophy*, 70: 556–67.

—— 1979. Counterfactual dependence and time's arrow. *Noûs*, Special Issue on Counterfactuals and Laws, 13: 455–76.

—— 1986. *Philosophical Papers*, vol. ii. New York: Oxford University Press.

Machamber, P., Darden, L., and Craver, C. F. 2000. Thinking about mechanisms. *Philosophy of Science*, 67: 1–25.

Mackie, J. L. 1965. Causes and conditions. *American Philosophical Quarterly*, 2: 245–64.

Marini, M. M., and Singer, B. 1988. Causality in the social sciences. *Sociological Methodology*, 18: 347–409.

Maudon, J., Malvin, J., Stiles, J., Nicosia, N., and Seto, E. 2000. Impact of California's Cal-Learn Demonstration Project: final report. UC DATA Archive and Technical Assistance.

Mellors, D. H. 1995. *The Facts of Causation*. London: Routledge.

Menzies, P., and Price, H. 1993. Causation as a secondary quality. *British Journal for the Philosophy of Science*, 44: 187–203.

Mill, J. S. 1888 *A System of Logic, Ratiocinative and Inductive*, 8th edn. New York: Harper and Brothers.

Neyman, J. 1990. On the application of probability theory to agricultural experiments: essay on principles, trans. D. M. Dabrowska and T. P. Speed. *Statistical Science*, 5: 463–80; first pub. in Polish 1923.

Papineau, D. 1985. Causal asymmetry. *British Journal for the Philosophy of Science*, 36: 273–89.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearson, K. 1911. *The Grammar of Science*, 3rd edn. rev. and enlarged, Part 1: *Physical*. London: Adam and Charles Black.

Pierson, P. 2004. *Politics in Time: History, Institutions, and Social Analysis*. Princeton, NJ: Princeton University Press.

Ragin, C. C. 1987. *The Comparative Method: Moving beyond Qualitiative and Quantitative Strategies*. Berkeley: University of California Press.

Riker, W. H. 1957. Events and situations. *Journal of Philosophy*, 54: 57–70.

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

—— 1978. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6: 34–58.

—— 1986. Statistics and casual inference: comment: which ifs have casual answers. *Journal of the American Statistical Association*, 81: 945–70.

—— 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5: 472–80.

Russell, B. 1918. On the notion of cause. In *Mysticism and Logic and Other Essays*. New York: Longmans, Green.

Salmon, W. C. 1990. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.

SEARLE, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. London: Cambridge University Press.

——1997. *The Construction of Social Reality*. New York: Free Press.

SHAFER, G. 1996. *The Art of Casual Conjecture*. Cambridge, Mass.: MIT Press.

SIMON, H. A. 1952. On the definition of the causal relation. *Journal of Philosophy*, 49: 517–28.

——and IWASAKI, Y. 1988. Causal ordering, comparative statics, and near decomposability. *Journal of Econometrics*, 39: 149–73.

SOBEL, M. E. 1995. Causal inference in the social and behavioral sciences. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, ed. G. Arminger, C. C. Clogg, and M. E. Sobel. New York: Plenum.

SORENSON, A. B. 1998. Theoretical mechanisms and the empirical study of social processes. In *Social Mechanisms*, ed. P. Hedström and R. Swedberg. Cambridge: Cambridge University Press.

SOSA, E., and TOOLEY, M. 1993. *Causation*. Oxford: Oxford University Press.

SPELLMAN, B. A., and MANDEL, D. R. 1999. When possibility informs reality: counterfactual thinking as a cue to causality. *Current Directions in Psychological Science*, 8: 120–3.

SPRINZAK, E. 1972. Weber's thesis as an historical explanation. *History and Theory*, 11: 294–320.

TETLOCK, P. E., and BELKIN, A. (eds.) 1996. *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton, NJ: Princeton University Press.

TILLY, C. 1984. *Big Structures, Large Processes, Huge Comparison*. New York: Russell Sage Foundation.

VAN FRAASSEN, B. 1980. *The Scientific Image*. Oxford: Clarendon Press.

VON WRIGHT, G. H. 1971. *Explanation and Understanding*. Ithaca, NY: Cornell University Press.

——1974. *Causality and Determinism*. New York: Columbia University Press.

WAND, J. N., SHOTTS, K. W., SEKHON, J. S., MEBANE, W. R., HERRON, M. C., and BRADY, H. E. 1991. The butterfly did it: the aberrant vote for Buchanan in Palm Beach County, Florida. *American Political Science Review*, 95: 793–810.

WAWRO, G. 1996. *The Austro-Prussian War: Austria's War with Prussia and Italy in 1866*. New York: Cambridge University Press.

WEBER, M. 1906 [1978]. *Selections in Translation*, ed. W. G. Runciman, trans. E. Matthews. Cambridge: Cambridge University Press.

WENDT, A. 1999. *Social Theory of International Politics*. Cambridge: Cambridge University Press.

CHAPTER 11

THE NEYMAN–RUBIN MODEL OF CAUSAL INFERENCE AND ESTIMATION VIA MATCHING METHODS

JASJEET S. SEKHON

"Correlation does not imply causation" is one of the most repeated mantras in the social sciences, but its full implications are sobering and often ignored. The Neyman–Rubin model of causal inference helps to clarify some of the issues which arise. In this chapter, the model is briefly described, and some consequences of the model are outlined for both quantitative and qualitative research. The model has radical implications for work in the social sciences given current practices. Matching methods, which are usually motivated by the Neyman–Rubin model, are reviewed and their properties discussed. For example, applied researchers are often surprised to

I thank Henry Brady, David Collier, and Rocío Titiunik for valuable comments on earlier drafts, and David Freedman, Walter R. Mebane, Jr., Donald Rubin, and Jonathan N. Wand for many valuable discussions on these topics. All errors are my responsibility.