Causal Inference in Educational Policy Research

David Kaplan
Department of Educational Psychology
University of Wisconsin – Madison

Introduction

With the passage of No Child Left Behind (NCLB), attention has focused on the need for evidenced-based educational research, particularly educational policies and interventions that rest on what NCLB refers to as "scientifically based research". In practice, this focus on scientifically based educational research has translated into a preference for research studies based on the principles of randomized experimental designs. Indeed, Part A., Sec. 9101 of the No Child Left Behind Act, under the definition "Scientifically Based Research" states

> "The term 'scientifically based research' (A) means research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs; and (B) includes research that ... (iv) is evaluated using experimental or quasi-experimental designs in which individuals, entities, programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition controls;..."

The ostensible reason for stating a preference for "random-assignment experiments" is the argument that the underlying theory of randomization provides a sound foundation for unambiguous causal inferences. The preference for randomized experimental designs stated in NCLB suggests that it is the "gold standard" for educational research. In fact, the standards of the U.S. Department of Education's Institute for Educational Sciences (IES) sponsored "What Works Clearinghouse, brooks no room for non-experimental studies as meeting the standard. From the standards document,

> Studies that provide strong evidence for an interventions effectiveness are characterized as *Meet Evidence Standards.* Studies that offer weaker evidence *Meet*

I am grateful to Jane Cooley, Adam Gamoran, Daniel Hausman, and Salvador Navarro for valuable comments and criticisms on an earlier draft of this chapter. The opinions expressed herein are mine alone.

THIS IS A DRAFT VERSION UNDER REVIEW. PLEASE DO NOT CITE OR CIRCULATE.

> *Evidence Standards with Reservations.* Studies that provide insufficient evidence *Does Not Meet Evidence Screens.* In order to meet evidence standards (either with or without reservations), a study has to be a randomized controlled trial or a quasi-experiment with one of the following three designs: quasi-experiment with equating, regression discontinuity designs, or single-case designs. (pg. 1, italics theirs).

Also, in a 2006 report prepared for the IES by the Coalition for Evidenced-Based Policy, it is clearly stated that "Well-designed and implemented randomized controlled trials are considered the "gold standard" for evaluating an interventions effectiveness, in fields such as medicine, welfare and employment policy, and psychology." (pg. 1, emphasis, theirs). For the purposes of this paper, those who advocate the *experimental* approach to policy analysis will be referred to as *experimentalists*

Although the methodology of randomized experimental designs can provide a strong basis for evaluating causal claims, it does not preclude the possibility that sound causal inferences can be drawn from non-experimental/observational settings. The choice of experimental or non-experimental designs is dependent on a number of factors and should not be based on an ill-defined notion of "scientifically-based research" or the presumption that any one methodology of inquiry constitutes the "gold-standard" of quality. In the context of causal inferences, there are strengths and weaknesses to both approaches. As will be described below, the randomized experimental design approach has the advantage of testing a well defined, albeit simple, counterfactual claim. However, experimental designs are typically employed to address relatively simple problems that fit a clinical trials-type model. In what will follow, I will argue that randomized experimental designs (a) are not suited to providing insights into the complexities of the educational system, (b)they are not structured to unpack the specifics of the treatment mechanisms operating as causal factors, (c) they do not provide an approach for testing numerous and more realistic counterfactual propositions, and finally (d) their efficacy in ruling out potential confounds can only be guaranteed in infinitely large samples.

In contrast to the experimental design approach, there exists a non-experimental/observational approach to causal inference that is grounded in classical macroeconomics and the method of simultaneous equation modeling. The econometric approach has had a long history in educational, psychological, and sociological research under the names *path analysis* and *structural equation modeling.* I argue in this paper that the econometric approach is suitable for complex problems in educational policy insofar as (a) it is suited to providing insights into the complexities of the educational system (b) the relative strength of counterfactual claims can be tested in concert with other counterfactual claims, (c) econometric models can be extended to handle unobserved heterogeneity, providing a way to examine how causal hypotheses operate in unobserved sub-populations, (d) multiple counterfactual conditional hypotheses can be examined, and (e) the approach doesn't rely on the hope of successful random assignment, and (f) it is readily suited for forecasting the effects of policies and interventions out-of-sample. These issues will be discussed in more detail below. However, for the econometric approach to be successful in the domain of educational research, a deeper synthesis of modern work on causal inference is required. Those who advocate the econometric approach to policy analysis will be referred to as *structuralists.*

The problem of estimating causal effects in educational research is of utmost importance, as we can all agree that it is crucial to provide a sound evidence base on which to test educational policies and interventions. Indeed, a recent monograph by Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson (2007) dealt explicitly with this issue by comparing and contrasting the experimental design approach with approaches to causal inference gleaned from observational and non-experimental designs in education. The Schneider et al (2007) monograph is to be commended for its emphasis on supporting the development of large scale educational databases to be used to test causal claims in education (a topic I will return to later in this chapter). However, the monograph did not provide an updated view of the philosophical or methodological developments relevant to a full understanding of how causal inference can be warranted using large scale databases. Their monograph does not cite philosophical work on counterfactual theory, nor does it provide a discussion of the major debates surrounding the limitations of the treatment effects approach in contrast to modern thinking relevant to a structural approach to educational research. Thus, the purpose of this chapter is to provide a review of modern philosophical and methodological positions on causal inference, and to synthesize these positions in the context of structural empirical educational research.

At the outset it must be noted that the topic of causation and causal inference is enormous and space limitations preclude a comprehensive literature review and synthesis. Therefore, the literature examined in this chapter is highly selective and does not represent all of the scholarship on the problem. With this caveat in mind, the chapter will begin with a brief overview of David Hume's notions of causation as it is Hume's writings that permeate much of the relevant theoretical and practical work that followed. The chapter then examines the standard critique of Hume. Although there are serious critiques of Hume's position, his writings set the groundwork for the experimental design tradition as embodied in the work of J. S. Mill, D. Campbell and J. Stanley, and more recently in the Donald Rubin and Paul Holland's model of causal inference. The experimental design tradition is then elucidated. A critique of the experimental design approach by Worrall (2002, 2004) is then discussed. The chapter then reviews somewhat more modern conceptions of causation, particularly the work of Mackie (1980) and Woodward (2003). I focus mainly on Mackie because, as I will argue, his work on counterfactuals serves to bridge the divide between the experimental design tradition and the econometric tradition. The chapter then proceeds to a discussion of causal inference from the econometric tradition embodied in the early work of Haavelmo, and more recently in positions put forth by Hoover (1990, 2001) and Heckman (2000, 2005). Focus in this section is on the idea of strengthening causal inferences in the context of observational studies, i.e. examining the causes of effects so as to provide more accurate predictions derived from policy counterfactuals. Next, I advance an argument that the structural approach is better suited to a science and practice of educational policy analysis than the experimental approach. I adopt the manipulationist view of causal analysis (Woodward, 2003) embedded the econometric approach but also bringing in Mackie's earlier ideas on counterfactual propositions. The chapter closes with a general conclusion outlining some of the important work that is missing from this review.

Early Historical Foundation

Providing a review of the historical foundations of the problem of causation is a difficult task. One can start by examining work dating far back in recorded history - at least as far back as the ancient Greeks and specifically Aristotle. However, for the purposes of a chapter that attempts to review the problem of causal inference as it relates to education policy analysis, it is convenient to start relatively more recently with the work of David Hume. The rationale for beginning with Hume is two-fold. First, As discussed in Hoover (2001) Hume's work on problems of causation stemmed from problems of economic policy analysis and therefore may be relevant to educational policy analysis. Indeed, Hume made very important contributions to theory of money as it was understood in the 18^{th} century. Second, much subsequent work on problems of causal inference can trace their origins to Hume's essential insights.

Hume's Philosophy of Causation

Hume's position on causation is contained in two of his seminal books: *A Treatise of Human Nature* (*THN*, 1739, Book I, Part III) and *An Enquiry Concerning Human Understanding* (*EHU*, 1777, Sections IV-VIII). To begin, in line with the philosophy of empiricism of his time, Hume viewed all human perception as arising from sense impressions. Intellectual ideas were no different and were also considered by Hume to be based ultimately in sense impressions, but perhaps of a more ephemeral sort. Thus, for Hume the question arises as to what constitutes the sense impressions that give rise to the perception of a necessary connection between events.

For Hume, the outward sense impression that leads to the perception of necessary connection is one of the *constant conjunction* of events. Using the famous metaphor of billiard balls colliding with one another, Hume argued that three elements are required to give rise to an inward idea of cause and effect. The first requirement is spatial contiguity, with one ball moving after it is touched by another. The second requirement is that the cause precedes the effect. The third requirement is *necessary connection* in the sense that the cause must reliably give rise to the effect.

In Hume's analysis, the first two requirements reside in the outward sense experience, while the last requirement, the inner idea of necessary connection, arises in the mind. For Hume, the idea of a necessary connection arising from the empirical experiences of contiguity and constant conjunction gives rise to the inner idea of necessary connection.

It appears that Hume is drawing the negative conclusion that we can have no outward knowledge of necessary connection or, for that matter, the forces of nature at all. This view is consistent with Hume's classic critique of induction which he articulates in *THN*. Hume writes, "There can be no *demonstrative* arguments to prove that *those instances, of which we have had no experience, resemble those, of which we have had experience*" (*THN* Book 1, Part 3, Sect. 6, italics Hume's)". For example, when an individual first experiences a billiard ball colliding with another and causing the second ball to move, that individual cannot logically form a general rule about future events. However, upon witnessing numerous instances of the same conjoined events, that individual will feel compelled to claim that they are necessarily connected, with one billiard ball moving, referred to as the *cause* and the other the moving upon being struck, referred to as the *effect*. Thus, necessary connection comes from the experience of numerous similar instances of conjunction. Hume argued though that this inductive fallacy is not mitigated by experiencing numerous like events any more than the experience of a single event. The only difference is that the experience of numerous instances of the event (unlike the single instance) leads the individual to *feel* that they are connected, and that now forecasting the future events is justified. This feeling of necessary connection is a habit or "custom" of the mind Hume (1739).

At first glance, it appears that Hume was espousing an anti-realist position. However, as pointed out by Hoover (2001), Hume was a realist in the sense that he believed that there were indeed actual forces of nature independent of our senses that cause events to take place. But, Hume denied that necessary connection is independent of our senses. Returning to the metaphor of the billiard ball, Hume accepted that there were natural forces at work that caused a billiard ball to move upon being struck by another billiard ball - Hume simply believed we could not know what those forces were. From Hume, "[W]e are never able, in a single instance, to discover any power or necessary connexion; any quality, which binds the effect to the cause, and renders the one an infallible consequence of the other" (*EHU*, Section 7, Part 1, pg. 136)".

Despite this rather negative conclusion, Hume argued that there was virtually nothing more important than the understanding of cause and effect. Hume wrote "By means of it alone we attain any assurance concerning objects, which are removed from the present testimony of our memory and senses". He further stated "The only immediate utility of all sciences, is to teach us, how to controul and regulate future events by their causes". Given Hume's view that a causal connection arises in the mind as a function of observing the constant conjunction of events, he arrived at his definition of causation - viz

> "[W]e 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 second.* Put another way *"...where, if the first object had not been, the second never had existed"* (Italics, Hume's).

What is particularly noteworthy about this definition is the second part. What appeared to Hume as a synonymous rephrasing of his definition of causation is, in fact, two different ways of considering cause. The latter part, which is of major concern in this chapter, concerns the role of the counterfactual conditional proposition in a theory of causation. The role of the counterfactual conditional proposition figures predominantly in later theories of causation and will be a major focus of this chapter.

Critiques of Hume

Hume's analysis of the problem of causation has been criticized on numerous grounds and was nicely summarized by Mackie (1980). For Mackie, perhaps the most serious problem relates to the condition of constant conjunction. Specifically, the argument that necessary connection arises in the mind after experiencing the constant conjunction of events, implies that constant conjunction is, indeed, constant - in other words, that the conjunction of events will always take place. But, by Hume's own critique of induction, this assumption is not tenable.

On the same point, Mackie (1980) noted that not all experiences of constant conjunction can sensibly give rise to the idea of necessary connection. For example, Mackie asked "...are there not causal sequences which nevertheless are not regular?" (1980, pg. 4). Mackie then answered his own question by giving an example of a coin toss. A coin toss is an indeterministic process, whereby each time I toss the coin, either it will land heads or land tails. Thus, on the one hand, tossing a coin is not a regular process, but on the other hand, surely my tossing the coin caused it to land heads. Mackie (1980) even questioned the necessity of assuming temporal precedence. He asked if there were not causes that could occur simultaneously with events. This latter concern is particularly crucial in the context of economic modeling as statistical models for supply and demand assume simultaneous causation. (However, see Fisher, xxxx, for a discussion of the some of the temporal assumptions associated with simultaneous equation modeling).

More recently, Pearl (2000) pointed to three problems with Hume's definition of causation. First, regularity, or correlation is simply not sufficient for causation. Second, Pearl argues that it is too strong to suggest that Hume's second aspect of causation - namely counterfactual sentences, is comparable to regular succession. That is, regular succession is based on observation and counterfactuals are mental exercises (Pearl, 2000, pg. 238). Third, Hume's addition of counterfactuals came nine years after his original definition that argued only for regularity. The addition of stating a counterfactual condition for causation, as I noted earlier, is not synonymous with the notion of regularity as stated in the first part of his second definition nor his earlier regularity definition. On the basis of these problems, Mackie (1980) concluded that Hume's definition of causation is "imprecise and carelessly formulated" (pg. 6).

The Experimental Design Tradition

Much more can be said regarding the problems with Hume's analysis of causation. For now, however, it is important to examine Hume's role in setting the stage for the problem of drawing causal inferences in experimental designs. In the context of educational policy analysis, this discussion is crucial because, as it was noted, the policies and practices of educational research under NCLB clearly articulates a preference for randomized designs in educational research. In this section, I discuss three important contributions to the underlying theory of experimental designs and the conditions that give such designs their efficacy. The first contribution is that of John Stuart Mill who, following Hume, set down many of the conditions of experimental designs. Indeed, Mackie (1980) notes that Mill's work was a vast improvement on Hume insofar as Mill recognized the existence of what Mackie much later referred to as *factors* that could operate as conditions for causation. The second contribution is that of Campbell and Stanley whose work on elucidating the confounds to causal conclusions drawn from experimental designs is considered the classic treatment on the subject. I also briefly examine extensions of Campbell and Stanley, including the work of Cook & Campbell. The third contribution is that of Rubin (1974) and Holland (1986), who provided the statistical underpinnings that give experimental designs their legitimacy in testing causal claims.

John Stuart Mill

In discussing the origins of the experimental design tradition, it is safe to start with ideas of Mill (1851). It should be noted, though, that many of Mill's ideas stemmed from some of Hume's basic thoughts about causation. Specifically, we find instances of Hume's contributions in Mill's different methods of experimental inquiry as espoused in his *System of Logic*. Nevertheless, Mill's ideas are fundamental to more modern treatments of experimental logic and design.

The goal for Mill was to isolate from the circumstances that precede or follow a phenomenon those that are linked to the phenomenon by some constant law. Mill (1851, System, III.viii.1). That is, the approach is to test if a presumed causal connection exists by experiment - by observing the relevant phenomena under a variety of experimental situations. To take an example from educational policy analysis, suppose we wish to know whether the reduction of class size causes improvement in achievement. How can that be proved? For Mill, causal inference rested on three factors. First, the cause must precede the effect; second, the cause and effect must be related; and third, other explanations for the cause-effect relationship must be ruled out. Mill's major contributions to experimental design concerned his work on the third factor for causal inference. Mill suggested three methods for dealing with this third factor. The first is the *Method of Agreement* which states that the effect will be present when the cause is present. The second is the *Method of Difference*, which states that the effect will be absent when the cause is absent. The third is the *Method of Concomitant Variation* which states that the given the first two methods, an inference of causation is stronger when other factors reasons for the covariation for the cause and effect are eliminated.

Campbell & Stanley

Perhaps the most important instantiation of Mill's codification of experimental logic, and the one that has had the most profound influence in experimental studies in the social and behavioral sciences is the work of citeA Campbell. In their small but seminal monograph *Experimental and Quasi-Experimental Designs for Research*, Campbell & Stanley lay out the logic of experimental and quasi-experimental designs. They provide the major sources of confounding in these types of designs and describe their strengths and weaknesses with regard to *internal* v. *external* validity.

For Campbell & Stanley, internal validity "is the basic minimum without which any experiment is uninterpretable" (pg. 5). In contrast to internal validity, the external validity of an experiment concerns the capability of the findings to generalize outside the experimental arrangement. These two types of validity often clash. Experimental designs that are strong in both internal and external validity are desirable; however, the fact remains that designs that are strong with respect to internal validity are often implemented at the cost of strong external validity. Indeed, Campbell & Stanley suggest that "While *internal validity* is the *sine qua non*, and while the question of *external validity*, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal" (pg 5, italics theirs).

A number of factors can serve as threats to the internal validity and hence causal conclusions that can be drawn from an experimental design. To take an example, consider the so called "one-group pretest-postest design". Here, measures are taken prior to the implementation of an intervention, after which the same measures are taken again. Campbell & Stanley provide this *bad example* of an experiment in order to elucidate the types of confounds that threaten causal claims that the intervention shifted the measures from the first to second observation. For example, the confound of *history* suggests that exogenous factors taking place between the first and second observation periods could be the explanation for the shift in means rather than the intervention. A second threat to the internal validity of the experiment is *maturation* where endogenous changes to the individual might induce a shift in the measures over time and that has nothing to do with the treatment intervention. These two threats, and many others, provide a sense of the types of criticisms that can be leveled against certain types of experimental arrangements.

To mitigate against these threats, Campbell & Stanley suggest adding a group that does not receive the intervention - the so-called *control group.* The addition of the control group along with the treatment group affords a very powerful defense against the threats to internal validity. The process by which the addition of a control group eliminates threats to the internal validity of an experiment is the concept of *random assignment.* Random assignment in this context simply means that each individual in the defined sample has an equal probability of being in the treatment group or the control group. With random assignment, all threats to internal validity as discussed in Campbell & Stanley are removed. Thus, for example in the case of maturation, we would expect that individuals in the treatment group are as likely to experience the same endogenous changes as those in the control group. It is the power of random assignment that lends experimental designs their strength. However, it is important to point out that random assignment removes threats to internal validity in infinitely large samples. In finite samples, problematic outcomes of random assignment to treatment and control can still occur, and these problems become more likely with smaller samples.

But what if random assignment is not feasible? In the nomenclature of Campbell & Stanley, designs that resemble true experiments in their arrangement but do not enjoy the benefits of random assignment are considered *quasi-experimental* designs. The ability to draw causal conclusions in the context of quasi-experimental designs is directly related to the degree of pre-treatment equivalence of the treatment and control groups. Numerous design and analysis methods have been employed to attempt pre-treatment equivalence. These include propensity score matching, covariate analysis, and others (see e.g ?, ?, ?). In the end, if pre-treatment equivalence can be attained, then most (but not all) of the threats to internal validity are addressed.

Cook & Campbell

The primary focus of Campbell & Stanley's monograph was to outline the conditions under which experimental arrangements were more or less robust to threats to internal validity. In a later and equally important work, Cook & Campbell extended the Campbell & Stanley framework to consider quasi-experimental designs that are applicable to field settings where random assignment might not be feasible. They begin their book by stating that "... this book is to outline the experimental approach to causal research..." and "... Since this book is largely about drawing causal inferences in field research, we obviously need to define cause" (pg. 1).

What follows in the first chapter of Cook & Campbell is an excellent overview of positions on the problem of causation, ending in what they refer to as an *evolutionary critical-realist perspective.* The evolutionary critical realist perspective of Cook & Campbell arises first out of the activity theory of causation proposed by Collingwood (1940) among others, and might now be referred to as the *manipulationist* view proposed by Woodward (2003). The idea is that what constitutes a cause is something that can be manipulated to bring about an effect. The evolutionary perspective refers to the notion that the human species (and perhaps other species) have a strong psychological predisposition to infer causal relations, and that this predisposition is the result of biological evolution and the survival value conferred on having such a predisposition. Their critical-realist perspective suggests that they view causes as actually operating in the real world but that our knowledge of these causes is conjectural and open to critical discussion. The critical nature of their work is in line with the critical-rationalist perspective of Karl Popper (xxxx) and his students.

The Rubin - Holland Model

A very important set of papers that have provided the statistical underpinnings for causal inference in experimental studies is the work of Rubin (1974) and Holland (1986) - here referred to as the Rubin - Holland Model. Their papers provide a framework for how statistical models that test causal claims are different from those that test associational claims, and that statistical theory has a great deal to add to the discussion of causal inference.

In the more recent of the two papers, Holland (1986) makes clear that his interest is in *"measuring the effects of causes* because this seems to be a place where statistics, which is concerned with measurement, has contributions to make" (pg. 945, italics Holland's). Holland is clear, however, that experiments are not the only setting where causal claims can be tested, but he does believe they are the simplest.

In outlining the Rubin-Holland model it is noted that their terminology of cause is not confined to cases of randomized experiments. The notion of cause (or, interchangeably *treatment*) in the Rubin-Holland model is relative to some other cause. Specifically, in considering the phrase *"X causes Y"*, the idea is that X causes Y relative to another cause - including the possibility of *"not X"*. Holland (1986) states that "For causal inference, it is critical that each unit must be potentially exposable to any one of its causes". Note how the Rubin-Holland model equates exposability to the notion of a counterfactual proposition. For example, in studies of class size and achievement, we can envision that the class size that a student is exposed to causes his/her achievement because we can envision exposing the same student to other class sizes. That is, we can set up a sensible counterfactual conditional statement of the sort "what if the student was not exposed to a change in class size". Rubin and Holland thus link exposability to counterfactual propositions.

To formalize these ideas, Holland starts by defining a selection variable S that assigns a unit u (e.g. an individual) who is a member of population U to either a treatment, t or a control c. In randomized experiments, S is created by the experimenter, but in observational (i.e. uncontrolled studies) such assignments often occur naturally. In the Rubin-Holland model, the critical characteristic is that the value *S(u)* for each individual could potentially be different.

The role of the outcome variable Y in the Rubin-Holland model is also crucial to their framework. First, for the variable Y to measure the effect of the cause, Y must be measured post-exposure - that is after exposure to the treatment [1] Then, the value of the

[1] The notion of temporal priority will be seen not to be necessary for causal inference.

post-exposure outcome variable must be a result of either the cause t or the cause c defined on particular unit. Therefore, the Rubin-Holland model conceives of the same individual providing an outcome variable after being exposed to the treatment, $Y_t(u)$ or after being exposed to the control $Y_c(u)$. The causal effect defined within the Rubin-Holland framework is then the difference between Y_t and Y_c for unit u. That is

$$Y_t - Y_c \tag{1}$$

This is the fundamental idea of the Rubin-Holland model - namely that causal inference is defined on individual units. However, as Holland (1986) points out, this fundamental notion has a serious problem - namely, that it is impossible to observe the value of Y_t and Y_c on the same unit and therefore impossible to observe the effect of t on u. Holland refers to this as the *Fundamental Problem of Causal Inference*.

Holland's approach to this fundamental problem is to draw a distinction between a *scientific solution* to the problem and a *statistical solution*. For Holland, the scientific solution requires that certain assumptions be made regarding temporal stability, causal transience, and unit homogeneity. Temporal stability refers to the assumption that the application of the control condition to a unit does not depend on when it occurred. Causal transience refers to the assumption that the response to the treatment is not affected by exposure of the units to the control condition - that the effect of the control condition is transient. Finally, unit homogeneity refers to the assumption that the response to a treatment applied to one unit is equal to the response to the treatment for another unit - that is $Y_t(u_1) = Y_t(u_2)$. As Holland notes, however, these assumptions are generally un-testable but are not uncommon in supporting causal claims in laboratory science.

The statistical solution to the Fundamental Problem offered by Holland (1986) is to make use of the population of individuals U. In this case, the *average causal effect*, T of t can be defined (relative to the control group) as the expected value of the difference between Y_t and Y_c over the units in the population - viz.

$$E(Y_t - Y_c) = T, \tag{2}$$

where T is the average causal effect, simplified as

$$T = E(Y_t) - E(Y_c). \tag{3}$$

To quote Holland (1986),"The important point is that the statistical solution replaces the impossible-to-observe causal effect of t on a specific unit with the possible-to-estimate average causal effect of t over a population of units" (pg 947. Italics Holland's)".

Much more can be said about the Rubin-Holland model, but what must be discussed is Holland's notion of what constitutes a cause, as his views are central to the arguments made in this chapter. Holland writes

> "Put as bluntly and as contentiously as possible... I take the position that causes are only those things that could, in principle, be treatments in experiments. The qualification, "in principle" is important because practical, ethical, and other considerations might make some experiments infeasible, that is, limit us to contemplating *hypothetical experiments*".

Holland goes on to say that the idea of what constitutes a cause is the same in both experimental and observational studies, except that in experimental studies, the investigator has a greater degree of control over the outcome than in the case of observational studies. From this, Holland points out that certain variables simply cannot be causes. For example, an attribute of an individual, such as gender or race cannot be a cause since the notion of *potential exposability* of the treatment is not possible without changing the individual. We cannot conceive of a situation in which we wish to know what an achievement score would be if a female child were male, because potential exposability is simply not possible. In the context of attributes, all that can be derived are associations, and although associations are important and suggestive of variables that might moderate causal variables, they cannot be causes in the sense of the Rubin-Holland model.

In the final analysis, four points are crucial to an understanding of the Rubin-Holland framework. First, the goal should be to seek out the effect of causes and not necessarily the causes of effects. For Holland, seeking out the causes of effects is valuable, but because our knowledge of causes is provisional, it is more valuable for a theory of causation to examine effects of causes. Second, effects of causes are always relative to other causes - particularly, the control. For Holland, and Campbell and Stanley before him, experiments that do not have a control condition are not experiments. Third, not everything can be a cause, and specifically, attributes cannot be causes. The law of causality simply states, according to Holland, that everything has a cause, but not everything can be a cause. Finally, for Rubin (1974) and Holland (1986), there can be *"no causation without manipulation"* (Holland, 1986, pg. 959).

Worrall's Critique of Randomization

A general argument that I make in this chapter is that randomized experiments do not represent a gold-standard of scientifically based research. This is not to say that randomized experiments are not a powerful approach to ascertaining a specific causal question, but it is not deserving of the approbation of the gold standard for research. A recent critique of randomized experiments by Worrall (2004, see also ; Worrall, 2002) supports my general view.

Worrall begins by examining the claim that randomization controls for selection bias - a bias that is arguably of most relevance to educational research. Worrall's argument is that this control is achieved not through the power of random assignment as obtained from, say, a coin toss, but rather from the experimenter not knowing in advance of the implementation of the treatment which group the unit (student, teacher, school, etc.) has been assigned to. Should an experimenter decide to remove, or otherwise change a unit's assignment after the coin toss has been made, then there is strong reason to suspect selection bias regardless of the initial randomization. In the context of educational research, it is worth asking if field experiments are actually blind, or more importantly, double blind. Moreover, as I will discuss later, there may be very good reasons to understand why units select in and out of treatments. Such information may be essential to understand in the context of scaling up an intervention.

Worrall continues by noting that randomization is not a guarantee of scientific validity nor is it a sure-fire approach to obtaining causal knowledge. Any reasonable proponent of randomization and randomized experiments understands that randomization controls for possible alternative causes of the response only in the long-run, viz. the average after conducting an infinite number of experiments. Therefore, the power of randomization is probabilistic, and given that experiments most often are conducted once (particularly in education), there is no reason statistically or epistemologically to believe that a single experiment guarantees an insight into causality.

In the interest of space, I will not say much more about Worrall's critique. Suffice to say that he is not advocating dispensing with randomization. Indeed, he argues that randomization controls for a specific type of confounder - namely selection bias, and this control can be easily compromised in practical settings. Moreover, randomization cannot control for all possible confounders, and its appeal to a limiting-average control does not improve the situation. Finally, and of most importance to this chapter, Worrall warns about the bad press given to observational studies or historically controlled trials.[2] He argues that if these types of studies are carefully conducted, we can obtain information about possible alternative explanations for outcomes that are at least as valid as those obtaining from randomized experiments.

In the context of educational research, I maintain along with Schneider, et. al (2007) that the existence of very well designed large scale educational databases, such as the Early Childhood Longitudinal Study (NCES, 1998), the Program on International Student Assessment (OECD, 2006), or the Trends in International Mathematics and Science Study (IEA, xxxx) are the types of observational studies that should be continually supported, developed, and utilized to test causal propositions. To do so, however, requires a bridging of counterfactual theories in the experimental literature and the structural (econometric) literature. This is the subject of the next section.

Counterfactual Propositions and a Manipulability Theory of Causation

Before discussing the structural approach for causal inference, more detailed attention must be paid to philosophical ideas regarding counterfactual propositions that connect to both the experimental tradition and the structural traditions. In the interest of space I will focus on the work of Mackie (1980), as it is his work on counterfactual propositions that helps to bridge the gap between the two traditions of causal inference. For an additional detailed study of counterfactuals, see Lewis (1973).

Mackie and the INUS Condition

Earlier in this chapter mention was made of Mackie's criticism of Hume's ideas about causality. In this section, I briefly outline Mackie's important contribution to our understanding of causation, as developed in his seminal work *The Cement of the Universe* (1980). I concentrate on two specific aspects of Mackie's work on causation because his ideas appear in later econometric treatments of causal inference, and which I believe lay a strong logical groundwork for how to consider causal inference in educational policy analysis. The first aspect of Mackie's work addresses a regularity theory of causation and the second aspect

[2] Interestingly, Worrall's critique is focused on medical research, not educational research where the press against observational studies is equally bad.

concerns a conditional analysis of causation. It should be understood that Mackie's overall contributions are much deeper than I have the space to present.

To begin, Mackie (1980) situates the issue of causation in the context of a modified form of a counterfactual conditional statement. Recall that a counterfactual conditional statement is of the form, *if X causes Y, then this means that X occurred and Y occurred, and Y would not have occurred if X had not.* This strict counterfactual conditional is problematic for the following reason; we can conceive of Y occurring if X had not. The example used by Mackie is that of striking a match. It is possible that a flame can appear without the act of striking the match, if, say, the match was lit by another source, for example another lit match. Thus, Mackie suggests that a counterfactual conditional must be augmented by considering the *circumstances* in which the causal event took place. In the case of the match, the circumstances under which striking the match produces the flame include no other potential cause of the match lighting. Thus, *under the circumstances*, the flame would not have occurred if the match had not been struck.

Mackie then discusses the distinction between *conditions* and *causes.* Another example used by Mackie is one of an individual lighting a match inside a flat of apartments, causing an explosion due to a gas leak. The temptation is to say that the explosion was caused by lighting the cigarette. In this regard, the gas leak is a standing condition and therefore not the cause of the explosion. However, as it is the case that lighting a cigarette in an apartment is not terribly unusual, but a gas leak is, we might be inclined to say that the gas leak was the cause of the explosion and not the lighting of the cigarette.

Mackie suggests that the problem in distinguishing between conditions and causes is addressed by considering that causes take place in a context, or what Mackie refers to as a *causal field.* In addressing the question of what caused the explosion, Mackie argues that the question should be rephrased as

> ”What made the difference between those times, or those cases, within a certain range, in which no such explosion occurred, and this case in which an explosion did occur?. Both cause and effect are seen as differences within a field; anything that is part of the assumed (but commonly understated) description of the field itself will, then, be automatically ruled out as a candidate for the role of cause”.

Mackie goes on to say

> ”What is said to be caused, then, is not just an event, but an event-in-a-certain-field, and some ’conditions’ can be set aside as not causing this-event-in-this-field simply because they are part of the chosen field, though if a different field were chosen, in other words if a different causal question were being asked, one of those conditions might well be said to cause this-event-in-that-other-field.” (pg. 35)

In the context of a causal field, there can be a host of *factors* that could qualify as causes of an event. Following Mackie (1980) let *A, B, C..., etc*, be a list of factors that lead to some effect whenever some conjunction of the factors occurs. A conjunction of events may be ABC or DEF or JKL, etc. This allows for the possibility that ABC might be a cause or DEF might be a cause, etc. So, all (ABC or DEF or JKL) are followed by the effect. For simplicity, assume the collection of factors is finite, that is only ABC, DEF, and JKL.

Now, this set of factors (ABC or DEF or JKL) is a condition that is both necessary and sufficient for the effect to occur. Each specific conjunction, such as ABC is sufficient but not necessary for the effect. In fact, following Mackie, ABC is a "minimal sufficient" condition insofar as none of its constituent parts are redundant. That is, AB is not sufficient for the effect, and A itself is neither a necessary nor sufficient condition for the effect. However, Mackie states that the single factor, in this case, A, is related to the effect in an important fashion - viz. [I]t is an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition: it will be convenient to call this ... an *inus* condition." (pg. 62)

It may be useful to briefly examine the importance of Mackie's work in the context of causal inference in educational policy analysis. Mackie's concept of *inus* conditions alerts us to the importance of carefully specifying the causal field in which causal claims are made, and to attempt to isolate those factors that serve as *inus* conditions for causal inferences.[3] Specifically, in the context of examining policies or interventions centered on improving reading proficiency in young children, Mackie would have us first specify the causal field or context under which the development of reading proficiency takes place. We could envision a large number of factors that could qualify as causes of reading proficiency. In Mackie's analysis, the important step would be to isolate the set of conjunctions, any one of which might be necessary and sufficient for improved reading proficiency. A specific conjunction might be phonemic awareness, parental support and involvement, teacher training. This set is the minimal sufficient condition for reading proficiency in that none of the constituent parts are redundant. Any two of these three factors is not sufficient for reading proficiency, and one alone - say focusing on phonemic awareness, is neither necessary nor sufficient. However, phonemic awareness is an *inus* condition for reading proficiency. That is, the emphasis on phonemic awareness is insufficient as it stands, but it is also a non-redundant part of a set of unnecessary but (minimally) sufficient conditions.

Mackie's analysis, therefore, provides a framework for unpacking the role of inus conditions for reading proficiency. I will argue later in this chapter that the econometric approach to causal inference provides the statistical grounding for examining Mackie's notion of the inus condition that is better suited to the context of educational policy analysis than the randomized design approach.

Woodward and the Manipulability Theory of Causation

A manipulability theory of causation was put forth by Woodward (2003). Briefly, Woodward first considers the difference between descriptive knowledge versus explanatory knowledge. While not demeaning the usefulness of description for purposes of classification and prediction, Woodward is clear that his focus is on *causal explanation.* For Woodward (2003), a causal explanation is an explanation that provides information for purposes of manipulation and control. To quote Woodward

> "... my idea is that one ought to be able to associate with any successful explanation a hypothetical or counterfactual experiment that shows us that and how [sic] manipulation of the factors mentioned in the explanation ... would be a way of manipulating or altering the phenomenon explained...Put in still another

[3] Pearl (2000) notes that in legal circles the convoluted meaning of the acronym *inus* has been replaced by the easier acronym *NESS* which stands for *"necessary element of a sufficient set".*

> way, an explanation ought to be such that it can be used to answer what I call the *what-if-things-had-been-different* question..." (pg. 11)

We clearly see the importance of the counterfactual proposition in the context of Woodward's manipulability theory. There is also a clear link to Holland's notion of potential exposability. Of equal importance is the invariance assumption described earlier. In fact, Woodward links invariance to interventions in a particularly important way. He argues that the relationship between the cause and effect must be invariant under an appropriate set of interventions. This is crucial for Woodward insofar as invariance under interventions distinguishes causal explanations from accidental associations.

It is certainly the case that the experimental approach allows one to ask the *what-if-things-had-been-different* question. This is the centerpiece of the Rubin-Holland framework because it builds this question at the level of the individual. However, the experimental approach does not provide a mechanism for studying the effect of hypothetical exposability within a well specified and theoretically grounded model. In line with Heckman, the treatment effects approach is more akin to "blind empiricism" (Heckman, 2005). The structural approach, while in agreement with the importance of manipulation and exposability, provides a broader and more realistic context for hypothetical manipulations. In the context of educational systems, many things can be different at one time, and these can be examined, in principle, in the structural framework. Nevertheless, although the statistical machinery for econometric modeling of educational data is in place, much better methods of large scale data collection are needed that allow for precisely stated counterfactuals within well developed models.

The Structural (Econometric) Tradition

In the previous two sections, I outlined the experimental approach, wherein the Rubin-Holland framework the notion of exposability and its relation to counterfactual propositions were discussed. I then reviewed Mackie's work on a modified counterfactual, bringing in his notion of the causal field and the inus condition. In this section I outline the structural approach as it relates to causal inference and in contrast to the experimental paradigm discussed earlier. I will focus attention on the methodology of simultaneous equation modeling as opposed to other standard statistical methodologies used in econometrics - such as time series analysis. Although issues of causality appear in the time series analysis literature, the simultaneous equations approach provides a richer framework for a discussion of causal inference and is also a methodology whose extensions have been applied in educational policy analysis under the names *path analysis* and *structural equation modeling* (see e.g. Kaplan, 2000).

This section begins with a very brief history of simultaneous equation modeling. It is recognized that this approach to economic models is not without criticism and the form of the criticism is briefly discussed. Regardless, I examine the issue of causal inference within this perspective and focus heavily on the work of Hoover (?) and Heckman (2000); ? (?).

Brief History of Simultaneous Equation Modeling

Mathematical models of economic phenomena have had a long history, beginning with Petty (1676, as cited in Spanos, 1986) However, the development of simultaneous equation modeling must be credited to the work of Haavelmo (1943). Haavelmo was interested modeling the interdependence among economic variables utilizing the form for systems of simultaneous equations.

The simultaneous equations model was a major innovation in econometric modeling. The development and refinement of the simultaneous equations model was the agenda of the Cowles Commission for Research in Economics, a conglomerate of statisticians and econometricians that met at the University of Chicago in 1945 and subsequently moved to Yale (see Berndt, 1991). This group wedded the newly developed simultaneous equations model with the method of maximum likelihood estimation and associated hypothesis testing methodologies (see Hood & Koopmans, 1953; Koopmans, 1950).

Formally, the simultaneous equation model can be written as

$$\mathbf{y} = \alpha + \mathbf{B}\mathbf{y} + \Gamma\mathbf{x} + \zeta, \tag{4}$$

where $\mathbf{y}$ is a vector containing the outcomes of interest - the "effects", $\mathbf{B}$ is a matrix of coefficients that allow the outcomes to be related to other endogenous outcomes and also allows for simultaneous relations among outcomes, $\mathbf{x}$ is a vector of exogenous variables - the "causes" that are measured, Γ is the matrix of coefficients that give the strength of the relationship between the causes and effects, and ζ is a matrix of structural disturbances which include unmeasured causes as well as random disturbances. For the next 25 years, the thrust of econometric research was devoted to the refinement of the simultaneous equations approach. Particularly notable during this period was the work of Fisher (1966) on model identification.

It is important to note that while the simultaneous equations framework enjoyed a long history of development and application, it was not without its detractors. As Heckman (2000) pointed out, by the mid-1960's the dominant view was that the program of the Cowles Commission was an "intellectual success but an empirical failure". Critics asserted that a serious problem with large macro-economic simultaneous equations models was that they could not compete with the relatively theory-free methods of the Box-Jenkins time series models and its multivariate extensions, so called vector auto-regressive (VAR) models, when it came to accurate predictions (e.g. Cooper, 1972). The underlying problem was related to the classic distinction between theory-based but relatively static models versus dynamic time-series models (see e.g. Spanos, 1986). The nature of the time-series approach and its widespread popularity was argued to be due to the fact that time-series models are more closely aligned with the data and therefore much better at prediction and forecasting than the more theoretical methods arising out of the structuralist approach to econometrics (Heckman, 2000). The counter argument to the time-series approach has been that it is not well suited for the evaluation of economic policy or testing policy relevant counterfactual claims.

Hoover and the Logic of Causal Inference in Econometrics

Within the structural tradition, an important paper that synthesized much of Mackie's notions of inus conditions for causation within the structural framework is the work of Hoover (1990); ? (?, ?). Hoover's essential point is that causal inference is a logical problem and not a problem whose solution is to be found within a statistical model per se. Moreover, Hoover argues that discussions of causal inference in econometrics are essential and that we should not eschew the discussion because it appears to border on realm of metaphysics. Rather, as with medicine, but perhaps without the same consequences, the success or failure of economic policy might very well hinge on a logical understanding of causation. A central thesis of this chapter is that such a logical understanding of causation is equally essential to rigorous educational policy analysis.

An important aspect of Hoover's work that is of relevance to our consideration of causal inference in educational policy is his focus on Mackie's inus condition. Adopting Hoover's notation, recall that the inus condition focuses on a set of antecedents A as a disjunction of minimally sufficient subsets of antecedent conditions for the consequence C. The comprehensive set A is the *full cause* of C whereas A_i is a *complete cause* of C. An element of A_i, say a_i is a cause of C if it is an insufficient but necessary member of an unnecessary but sufficient set of antecedents of the effect C. Thus, in line with Mackie's analysis, Hoover suggests that the requirement that a cause be necessary and sufficient is too strong, but, necessity is crucial in the sense that, as in line with Holland (1986), every consequence must have a cause.

Hoover sees the inus condition as particularly attractive to economists as it focuses attention on some aspect of the causal problem without having to be concerned directly with knowing every minimally sufficient subset of the full cause of E. In the context of education policy analysis, these ideas should also be particularly attractive. As in the example of reading proficiency used earlier, we know that it is not possible to enumerate the full cause of reading proficiency, but we may be able to focus on an inus condition - say parental involvement in reading activities.

From here, Hoover draws out the details of the inus condition specifically as it pertains to the structuralist perspective. Specifically, in considering a particular substantive problem, such as the causes of reading proficiency, we may divide the universe into antecedents that are relevant to reading proficiency, A and those that are irrelevant *non-A*. Among the relevant antecedents are those that we can divide into their disjuncts A_i and then further restrict our attention to the conjuncts of particular inus conditions. But what of the remaining relevant causes of reading proficiency in our example? According to Mackie, they are relegated to the *causal field*. Hoover considers the causal field as the standing conditions of the problem that are known not to change, or perhaps to be extremely stable for the purposes at hand. In Hoover's words, they represent the "boundary conditions" of the problem.

But the causal field is much more than simply the standing conditions of a particular problem. Indeed, from the standpoint of classical linear regression, those variables that are relegated to the causal field are part of what is typically referred to as the *error term*. Introducing random error into the discussion allows Mackie's notions to be possibly relevant to indeterministic problems such as those encountered in educational policy analysis. However, according to Hoover, this is only possible if the random error terms are components of Mackie's notion of a causal field.

Hoover argues that the notion of a causal field has to be expanded for Mackie's ideas to be relevant to indeterministic problems. In the first instance, certain parameters of a causal process may not, in fact, be constant. If parameters of a causal question were truly constant, then they can be relegated to the causal field. Parameters that are mostly stable over time can also be relegated to the causal field, but should they in fact change, the consequences for the problem at hand may be profound. In Hoover's analysis, these parameters are part of the boundary conditions of the problem. Hoover argues, most interventions are defined within certain, presumably constant, boundary conditions. In addition to parameters, there are also variables that are not of our immediate concern and thus part of the causal field. Random errors, in Hoover's analysis contain the variables omitted from the problem and "impounded" in the causal field.

> "The causal field is a background of standing conditions and, within the boundaries of validity claimed for the causal relation, must be invariant to exercises of controlling the consequent by means of the particular causal relation (INUS condition) of interest" (Hoover, 2000, pg. 222)

Hoover points out that for the inus condition to be a sophisticated approach to the problem of causal inference, the antecedents must truly be antecedent. Often this is done by appealing to temporal priority, but this is sometimes unsatisfactory. Hoover gives the example of laying one's head on a pillow and the resulting indentation in the pillow as an example of the problem of simultaneity and temporal priority [4]. Mackie, however, sees the issue somewhat more simply - namely the antecedent must be directly controllable. This notion of direct controllability, which is also an important feature of the Rubin-Holland model, leads to the problem of *invariance.* Invariance is essential to causal claims, and particularly counterfactual propositions. Hoover as well as Cartwright (1989) notes that an antecedent must have the capacity to change a consequent, and that capacity must be somewhat stable over time. The stability of a relationship in response to control of the antecedent is the problem of invariance. It is also related to the problem of causal ordering, as well be described later.

Employing an example by Hoover, but contextualized for educational research, consider as a true data generating mechanism[5] the following model for a sample of i children ($i = 1, 2, \ldots N$)

$$R_i = \beta P_i + \epsilon_i, \tag{5}$$

$$P_i = \mu + \zeta_i, \tag{6}$$

where in equation (5) R_i is a measure of reading achievement for student i, P_i is a measure of parental reading practices associated with student i, β is a regression coefficient relating the reading achievement to parental reading practices, and ϵ_i is a random disturbance term.[6] Equation (6) simply describes the marginal distribution of parental reading practices, where for this example it is assumed to be normal. [7]. The *reduced form* of the equation is obtained by inserting equation(6) into equation (5) and gathering terms. This yields

[4] This example was originally put forth by Kant in the context of an iron ball depressing a cushion

[5] The term *data generating mechanism* or *DGP* is used most commonly in the econometric literature and refers to the actual real-life process that generated the data. A statistical model ideally captures the true *DGP*.

[6] This model ignores the problem of nesting that is common in educational research. This is done for simplicity of the discussion and represents no loss of generality to the argument.

[7] The normality assumption is not trivial and relates to the statistical issue of *weak exogeneity*, which is a crucial issue for simulations of counterfactual propositions. see xxxx

$$R_i = \beta(\mu + \zeta_i) + \epsilon_i, \tag{7}$$
$$= \beta\mu + \beta\zeta_i + \epsilon_i, \tag{8}$$
$$P_i = \mu + \zeta_i. \tag{9}$$

From basic statistical theory, the joint distribution of reading achievement and parental reading practices can be factored into the conditional distribution of reading achievement given parental practices (the regression function) times the marginal distribution of reading achievement - viz.

$$f(R, P) = f(R|P)f(P) \tag{10}$$

But similarly,

$$f(R, P) = f(P|R)f(R) \tag{11}$$

Again, under the assumption that Equations (5) and (6) represent the true data generating model, then when these two forms are expressed in terms of model parameters, and assuming normality, we get

$$f(R|P) \sim N(\beta P, \sigma^2_\epsilon) \tag{12}$$
$$f(P) \sim N(\mu, \sigma^2_\zeta) \tag{13}$$
$$f(P|R) \sim N\left(\frac{\beta\sigma^2_\zeta P + \mu\sigma^2_\epsilon}{\beta^2\sigma^2_\zeta + \sigma^2_\epsilon}, \frac{\sigma^2_\epsilon\sigma^2_\zeta}{\beta^2\sigma^2_\zeta + \sigma^2_\epsilon}\right) \tag{14}$$
$$f(R) \sim N(\beta\mu, \beta^2\sigma^2_\zeta + \sigma^2_\epsilon) \tag{15}$$

What is important about this discussion is the influence of hypothetical or real changes in reading achievement or parental practices, when the true underlying data generating model is given Equations (5) and (6). Specifically, consider a program that is designed to increase parental reading practices. Then, this implies that the parameters of the marginal distribution of P, namely μ or σ^2_ζ will change. Noting that these parameters appear in the conditional distribution $f(P|R)$ they will be expected to change, and in the same vain, the marginal distribution of reading proficiency will change. However, the true conditional distribution $f(R|P)$ will remain invariant. Suppose instead that the "reverse regression" is wrongly specified and a policy or program is implemented to raise reading achievement without explicitly operating on parental practices. Then, from Equations (12) - (15) we see that the marginal distribution of reading proficiency $f(R)$ and conditional distribution of parental involvement given reading proficiency $f(P|R)$ will also change. In addition, though, the $f(R|P)$ will change. The conclusion that can be drawn from this example is that the true relationship between reading proficiency and parental involvement is invariant to policies and interventions related to parental involvement. When the wrong causal relationship is specified, invariance does not hold.

What has been described is the problem of *parameter invariance*, which is crucial for drawing causal inferences regarding policies and interventions and which, for Hoover, constitutes a possible strategy for determining causal order. Specifically, using historical information from large representative longitudinal or trend data, we may be able to locate periods of time in which no major interventions took place that would change reading achievement or parental reading practices. Then, either causal order would yield stable regression coefficients. However, during periods of time in which there have been policies that were targeted toward reading achievement and policies or interventions targeted toward parental reading practices, then examining the relative stability of the coefficients to the different factorizations would provide information regarding causal order.

Although this is a somewhat contrived example, it is offered to make two important points about the structural approach to causal inference in educational research. First, as discussed by Cartwright (1989) invariance is an essential foundation for counterfactual propositions. Using Cartwright's notation, the antecedent of a counterfactual proposition, say C, must have a stable causal relationship to the consequent, say E, when C comes under direct control. To quote Cartwright, "If Cs do ever succeed in causing Es (by virtue of being C), it must be because they have the capacity to do so. That capacity is something they can be expected to carry with them from situation to situation (pg. 145)". Second, the structural approach provides a way to examine causal orderings, and this alone can provide important insights into the functioning of complex systems such as education. Gaining an understanding of causal ordering and invariance is crucial for ascertaining the behavior of policies and interventions - particularly as those policies or interventions are being taken to scale. The structural approach can provide insights into these problems which are completely overlooked in the experimental approach.

Heckman's Scientific Model of Causality

Recently, Heckman (2000, 2005) provided important contribution to the problem of causal inference from a structural econometric perspective. Heckman's perspective centers on two essential points. First, Heckman views causality as the property of a model of hypothetical statements and that a fully developed model should represent a set of precisely formulated counterfactual propositions. According to Heckman, the problem with modeling the effects of causes, which is the mainstay of the experimental paradigm, is that such a perspective does not speak to how the process on which causal inferences are being drawn has been generated. In Heckman's words.

> "The ambiguity and controversy surrounding discussion of causal models are consequences of analysts wanting something for nothing: a definition of causality without a clearly articulated model of the phenomenon being described (i.e. a model of counterfactuals)". pg 2.

For Heckman, models are descriptions of hypothetical worlds and how these hypothetical worlds change as a function of manipulating the variables that determine the outcomes. Thus for Heckman, science is about constructing these causal models, and that the growth of human knowledge is based on constructing counterfactuals and developing supportive theories. In contrast, the experimental paradigm, now popular in education, represents a type of "blind empiricism" that, unguided by theory, will lead nowhere (Heckman, 2000).

Heckman's second point is that causal inference within statistics conflates three tasks that he argues need to be clearly separated: (a) definitions of counterfactuals, (b) identification of causal models from population distributions, and (c) identification of causal models from actual data. Heckman views the definition of counterfactuals as located in the realm of a scientific theory. The problem of the identification of causal models from population distributions falls into the purview of the mathematical analysis of identification (see Fisher). Finally, identification of causal models from actual data is a problem for estimation and sampling theory.

Heckman then compares his view with the approach to causal inference favored in epidemiology and clinical drug trials, and now educational research - namely the randomized experimental design approach described earlier. In Heckman's view, there are two essential problems with the experimental design approach. The first problem relates to the issue of selection bias - namely, the experimental approach does not model the mechanism by which counterfactuals are selected or how hypothetical interventions might be realized. This problem is seen when units making the choices are not the same as the units receiving the treatment - as in parents ultimately making schooling choices for the children although the latter may be those exposed to an intervention. The structural approach, guided by theory, can provide information that would allow the construction and testing of various selection mechanisms. The capability of structural models to provide insights into the selection process is potentially of great importance in education in the context of taking an intervention to scale.

The second problem is that the experimental approach does not specify the sources of randomness embedded in the error terms. Modeling these unobservable sources of random variation is, according to Heckman, essential in choosing the correct estimation method. Heckman goes on to suggest that the "treatment" in randomized designs is a conglomerate of factors that are not related to a theory of what actually produced the effect. Finally, and perhaps of most relevance to educational policy analysis, the experimental approach cannot be used for out-of-sample forecasting to new populations. In the context of educational policy analysis, the experimental approach does not yield insights into the behavior of the treatment when scaled up. The structural approach, on the other hand

> "...like the goal of all science, is to model phenomena at a deeper level, to understand the causes producing the effects so that we can use empirical versions of the models to forecast the effects of interventions never previously experienced, to calculate a variety of policy counterfactuals, and to use scientific theory to guide the choices of estimators and the interpretation of the evidence. These activities require development of a more elaborate theory than is envisioned in the current literature on causal inference in epidemiology and statistics." (Heckman, 2005)

Of relevance to the application of the Heckman's approach to educational policy, the structural approach allows historically experienced policies to be examined in light of new policies not experienced. Of course, in the context of educational policy analysis, this depends on support for research and development into better strategies for large scale data collection.

A third and quite serious problem is that the experimental approach cannot address the problem of general equilibrium. To take a simple example of the issue, consider the problem of class size reduction. In the context of an experimental design studying the effects of class size reduction on student achievement, the experimental approach would provide an estimate of the average causal effect of the treatment on the achievement outcome of interest. Other variables that would be affected by class size reduction and are correlated with achievement (e.g. teacher quality) cannot be addressed in the experimental framework, and indeed, in the context of the experiment are averaged out due to random assignment. However, the influence of these other variables becomes extremely serious with regard to the success of class size reduction when the policy is taken to scale. Thus, if a policy to reduce class size was, in fact, implemented (e.g. all middle school math classes would be reduced to no more than 15 children in a classroom) the effect on teacher hiring, teacher quality teacher salaries, buildings, and a host of other important variables, could all be effected. The adjustments that take place in these other variables represent the general equilibrium effects. Again, the experimental approach simply does not address these issues, whereas the structural approach can provide an explicit framework for modeling these effects and to test how various changes in the causal variable of interest manifest themselves in terms of the general equilibrium effects.

Toward an Approach to Causal Inference Suitable For Educational Policy Analysis

In this section, I argue that the structural approach to causal inference advocated by Heckman (2005) as well as Hoover (1990) is better suited to a science and practice of educational policy analysis than the experimental approach. First, it is widely accepted that educational systems are extremely complex, hierarchically organized systems of actors. The reality and consequences of this complexity for educational policy analysis is simply not captured by experimental designs, which focus on relatively narrow questions that align well with the model for clinical drug trials. Second, the experimental approach is not sufficiently detailed in unpacking the causal mechanisms responsible for any observed treatment effect, thus risking problems when going to scale with the treatment. Third, the experimental approach does not provide any information on how selection might operate in the choice of intervention or outcome. Fourth, the experimental approach does not provide a framework for examining ”out-of-sample” predictions based on varying and realistic counterfactual propositions. Finally, the experimental approach cannot address general equilibrium effects that would likely operate when an intervention or policy is taken to scale. As a result, and in line with Heckman (2005) and more generally Worrall (2002, 2004) the experimental approach makes too many implicit assumptions and presents too simplistic a view of the educational system to aid in building a knowledge base that can serve in developing effective interventions within a science of educational policy analysis.

In contrast, the structural approach advocated by Heckman (2005), particularly when supplemented by Hoover's synthesis of Mackie's inus conditions, contains the methodology for formulating a number of precisely stated counterfactuals within a well specified, albeit hypothetical model. The structural approach allows further testing of varying counterfactual propositions that can reflect the complex reality of educational systems. Such modeling efforts can provide important substantive feedback regarding how policies or interventions might work to effect outcomes under different counterfactual scenarios, including issues of treatment and/or outcome selection. In a related manner, the flexibility of the structural approach allows examining general equilibrium effects under a variety of realistic scenarios that could be faced when an intervention or policy goes to scale. The structural approach to educational policy analysis is much better suited to examine the potential for how policies and interventions might operate out-of-sample. Finally, the structural approach can mitigate against the potential the "blind empiricism" that Heckman (2005) feels characterizes the experimental approach, and can provide a framework for theory development in educational policy analysis.

The advantages of the structural approach to causal inferences notwithstanding, much more work needs to be done to fully integrate this approach into educational research. This is particularly true given the way in which path analysis and structural equation modeling have traditionally been applied to education, sociology, and psychology. A critique of the standard approach to path analysis and structural equation modeling as applied in education, psychology, and sociology is given in Kaplan (2000)). The essence of the critique is that conventional applications of structural equation modeling have not gone much beyond the presentation of goodness-of-fit measures. Although goodness-of-fit is important as it provides information regarding how well the model matches the data generating process, additional information can be gained from other forms of model evaluation. Interestingly, this view was recently expressed by Keane in a comparing the experimentalist approach to the structuralist approach to structural modeling in economics. Keane argued for much greater effort in testing model predictions "out-of-sample".

Also, much greater effort needs to be focused on precisely articulating identifying assumptions as they pertain to conceptual or theoretical frameworks. Specifically, a conceptual framework and the theoretical equations that are suggested by the framework do not necessarily imply that causal parameters can be uniquely identified from the data. In some cases, identifying restrictions must be imposed, and these restrictions require justification within the theoretical framework. Furthermore, in the context of cross-sectional data, it is essential that assumptions associated with the estimation of "contemporaneous" equations be carefully argued. For example, in a cross-sectional study, there is the implicit assumption that only the exogenous variables at the current time point are relevant to explaining the outcome, or that exogenous variables are unchanging and they capture the entire history of the inputs, and the exogenous variables are unrelated to any unobserved set of variables (Todd & Wolpin, 2003). All of these issues point to the need for bold theoretical development in educational policy analysis and aggressive support for research and development in high quality data sources that can be brought to bear on testing theoretical claims.

It might be argued that my critique of the experimental approach to education policy analysis is unfair - that this approach has always been intended to follow the clinical trials model so as to ascertain "what works" in education, and that it was never intended to go beyond that. This would be a fair criticism if it weren't for the unfortunate fact that there is a stated preference for randomized experimental designs in the language of No Child Left Behind. Indeed, the characterization of randomized experimental designs as the "gold-standard" for educational research have been found in other (U.S. Department of Education sponsored) writings and clearly implies that other methodologies of causal inference in education are relatively inferior - an implication that is both unfair, inaccurate, and unproductive. If the goal of this aspect of NCLB was simply to advocate for a methodology designed to ascertain a narrowly defined, albeit important, set of questions, then the rhetoric linking randomized experiments to the all-encompassing phrase "scientifically based research" would have been avoided, and other empirical methodologies would have been given equal preference. However, the NCLB Act clearly attempts a definition of ”scientifically based research” centered on a specific *methodology* of science and not a preference for multiple empirical *scientific* methodologies. Under NCLB, theory development and testing that incorporates broad methodological strategies does not constitute ”scientifically based research”. Instead, ”scientifically based research”, as defined in NCLB, constitutes a narrow utilitarian focus on the effects of causes utilizing a specifically preferred method for ascertaining ”what works”. And, although it is important to isolate evidenced-based interventions that can aid in ameliorating problems in education, the randomized experimental design is not the only, or necessarily the best, ”scientifically based” approach to finding solutions to the problems plaguing education in the United States.

The unfortunate rhetoric of NCLB aside, I also argue that neither the experimental or structural approach can legitimately stake a claim to being the gold standard for methodological rigor. In fact, there are overlapping areas of agreement with respect to the experimental and structural approaches and these areas of agreement are useful to exploit as we attempt to improve methodologies for causal inference in educational policy research. First, both approaches rightly reject the nihilistic post-modern relativism that seems to have infected education research of late. Second, both approaches urge rigorous standards of empirical data collection and measurement. Third, both approaches view causal relationships as having to satisfy counterfactual conditional propositions.

It is this last area of agreement where I believe there is considerable overlap between the Rubin-Holland treatment effects approach and the econometric approach advocated by Heckman, Hoover, and others. Perhaps the main area of agreement is the notion of exposability to potential treatment conditions and the general notion of manipulability articulated by Woodward (2003) as the basis for inferring causation. However, given their shared agreement in the importance of manipulability, potential exposability, and counterfactuals, it is surprising that neither Holland nor Heckman mention Mackie’s (1980) philosophical analysis of counterfactual propositions and inus conditions. Moreover, neither Holland nor Heckman refer to Hoover’s (1990; ?) work on causal analysis in econometrics. Thus, a fruitful area of research would be to link the structural approach to the experimental approach via Mackie’s ideas of counterfactuals, causal fields, and inus conditions.

It is also the case that the experimental approach and structural approach can be combined in fruitful ways. Indeed recent research has shown that experimental studies can be used to validate the predictions developed from complex structural models. In a recent paper, Todd and Wolpin (2006) examined the effect of a school subsidy program in Mexico that was implemented as a randomized social experiment. In addition, they specified a complex dynamic structural model that captured parental decisions about fertility and child schooling. The model was capable of reproducing the treatment effect quite closely. However, whereas the results of the social experiment stopped at the treatment effect, the structural model allowed specification and testing of other ex ante policy scenarios that produced information regarding general equilibrium effects. In the context of education policy, we can imagine a similar exercise in the context of, say, class-size studies. Specifically, we can envision attempting to estimate a model of class size, examine its prediction under conditions similar to the class size experiment in order to use the experimental results to validate the model, but then use the model to examine a variety of realistic policy alternatives. Clearly, we would need to anticipate the type of data needed to conduct such a study.

Conclusion

As noted in the introduction, a considerable amount of research on the problem of causal inference has been omitted from this chapter. There is simply not enough space to cover all of the work on this problem, and I have likely even omitted writings that are relevant to both a defense and criticism of my central arguments. Suffice to say that there are other essential ideas that deserve further exploration within the context of advancing methodologies for causal inference in educational policy research. For example, I did not thoroughly review the work on probabilistic causality, originally considered by Suppes (1970), and expanded by Eells (1991). My review also did not cover the important work of Pearl (2000) or the work of Spirtes, Glymour, and Scheines (2000). Space precluded a full discussion of Cartwright's notions of causes as capacities, nor have I discussed Hausman's (1998) work on causal asymmetry. The work of these writers, and many others, must be thoroughly examined as we consider building a rigorous science of causal inference for education policy research.

Space limitations notwithstanding, the goal of this chapter was to review certain central ideas of causal inference and to argue for an approach to causal inference in educational policy analysis that rests on philosophical and methodological work arising from macroeconomic policy modeling. The arguments raised in this chapter speak to the need to support basic and applied research on methodologies for non-experimental and observational studies and to vigorously support research and development into the design and analysis of large scale databases as a means of testing out invariance assumptions and hypothetical counterfactual experiments. In summary, although it is clear that much more philosophical and methodological work remains, it is hoped that this chapter will stimulate a broader discussion on the development of a rigorous empirical science and practice of educational policy analysis.

References

Cartwright, N. (1989). *Nature's capacities and their measurement.* Oxford: Oxford University Press.

Eells, E. (1991). *Probabilistic causality.* Cambridge: Cambridge University Press.

Hausman, D. M. (1998). *Causal asymmetries.* Cambridge: Cambridge University Press.

Heckman, J. J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, *115*, 45-97.

Heckman, J. J. (2005). The scientific model of causality. In R. M. Stolzenberg (Ed.), *Sociological methodology* (Vol. 35, p. 1-97). Boston: Blackwell Publishing.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945-960.

Hoover, K. D. (1990). The logic of causal inference: Econometrics and the conditional analysis of causality. *Economics and Philosophy*, *6*, 207-234.

Hoover, K. D. (2001). *Causality in macroeconomics.* Cambridge: Cambridge University Press.

Hume, D. (1739). *A treatise of human nature.* Oxford: Oxford University Press.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Newbury Park, CA: Sage Publications.

Mackie, J. L. (1980). *The cement of the universe: A study of causation.* Oxford: Oxford University Press.

Mill, J. S. (1851). *A system of logic* (3rd ed., Vol. I). London: John W. Parker.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge: Cambridge University Press.

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

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimation causal effects: Using experimental and observational designs.* Washington, DC: American Educational Research Association.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search.* Cambridge, MA: The MIT Press.

Suppes, P. (1970). A probabilistic theory of causality. *Acta Philosophica Fennica*, *24*, 5-130.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *2*, 3-33.

Todd, P. E., & Wolpin, K. I. (2006). Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *American Economic Review*, *96*, 1384-1417.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.

Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science*, *69*, S316-S330.

Worrall, J. (2004). *Why there's no cause to randomize* (Tech. Rep.). Centre for Philosophy of Natural and Social Science.