Essays, opinions, and professional judgments are welcome in this section of AJE. ***Forum*** *articles speak to and about the philosophical, ethical, and practical dilemmas of our profession. Authors are invited to express their views to encourage constructive dialogue centered on issues. To keep the dialogue substantive, other articles motivated by previous* ***Forum*** *presentations should have independent titles and themes. Items labeled "Comments on..." and "Rejoinder to..." will not be published in* ***Forum****—such responses are welcome and encouraged in the* ***In Response*** *section of AJE. Standard citations and reference lists should acknowledge and identify earlier contributions and viewpoints. Manuscripts should not exceed 10 double-spaced typewritten pages in length, unless the paper is invited by the editor.*

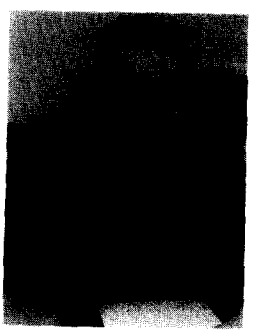
The views expressed in this section are independent of and not necessarily shared by AJE editors. By designs "Forums" are open to diverse views, in the hope that such diversity will enhance professional dialogue.

The Qualitative Method of Impact Analysis

LAWRENCE B. MOHR

ABSTRACT

Consider the qualitative approach to evaluation design (as opposed to measurement) to be typified by a case study with a sample of just one. Although there have certainly been elaborate and emphatic defenses of the qualitative approach to program evaluation, such defenses rarely attempt to qualify the approach explicitly and rigorously as a method of impact analysis. The present paper makes that attempt. The problem with seeking to advance a qualitative method of impact analysis is that impact is a matter of causation and a non-quantitative approach to design is apparently not well suited to the task of establishing causal relations. The root of the difficulty is located in the counterfactual definition of causality, which is our only broadly accepted formal definition of causality for social science. It is not, however, the only definition we use informally. Another definition, labeled "physical causality," is widely used in practice and has recently been formalized. Physical causality can be applied to the present problem. For example, it explains the persuasiveness of Scriven's "Modus Operandi" approach



Lawrence B. Mohr

Lawrence B. Mohr • SPP-Lorch Hall, University of Michigan, Ann Arbor, MI 48109-1220; Tel: 313-763-4588; E-Mail: lmohr@umich.edu.

American Journal of Evaluation, Vol. 20, No. 1, 1999, pp. 69–84.
ISSN: 1098-2140

to causal inference in evaluation. Under this conceptualization, a tailored case study design with a sample size of one becomes in principle as strong a basis for making inferences about program impact as a randomized experiment. Crucial for the application of the method to program evaluation is the finding that people's "operative reasons" for doing what they do are the physical causes of their actions. Lastly, it is shown that external validity using this qualitative approach would have exceptional strengths.

INTRODUCTION

Qualitative designs for program evaluation have become widely accepted and practiced (Greene, 1994; Goddard & Powell, 1994). The idea has also generated a good deal of controversy (summarized in part in Chen, 1990) wherein each side in the debate—qualitative vs. quantitative—finds its own method to be clearly superior and the other to be woefully inadequate. Many authorities take what they see as a middle ground, advocating that both approaches may be used in concert (Reichardt & Cook, 1979; Chen, 1990; Greene & Caracelli, 1997). It has begun to be increasingly clear, however, that although both approaches may well have something to contribute to program evaluation, their areas of contribution are not identical. Although there is undoubtedly overlap between the two approaches in what they seek to accomplish, they are not substitutes for one another.

Quantitative designs or approaches have been seen by their advocates as being especially good for impact analysis. Qualitative designs, on the other hand, have been advocated primarily in conjunction with other evaluative functions such as implementation analysis, process analysis, community self-analysis, the empowerment of staff, and the interpretation and understanding of experience. The functions in this latter list will not be reviewed here because they are not germane to the present purpose, which focuses on impact analysis. Maxwell and a small number of additional scholars (summarized in Maxwell, 1996) have begun to explore a role for qualitative work in the assessing of impacts. More typical, perhaps, the defense of the case study method by one of its most outspoken proponents in program evaluation does not credit it with efficacy in the assessment of impacts and distinguishes it from quantitative research in part on that basis (Stake, 1981). Of course, some proponents of the qualitative approach suggest that it is better than the quantitative approach precisely because its particular goals or functions have greater value than the goal of impact analysis (see, for example, Guba & Lincoln, 1981), but that debate is also outside of the present scope.

Let us assume, recognizing that some might deny it, that there often is some value to impact analysis in program evaluation. If one were concerned with impact analysis and if there were a qualitative approach that did indeed have clear relevance for that function, then it would be reasonable to bring it to light and discuss it. The purpose of this paper is to advance the idea of one such qualitative approach and to provide an introductory treatment of it. This idea must of necessity be viewed against a backdrop composed of concepts that, while familiar to the evaluator, are here combined in a nontraditional picture that will help bring clarity to the issue at hand.

CAUSATION

Impact analysis deals with causation. While some may take the position that program evaluation as a discipline should rid itself of the preoccupation with causation (or equivalently here,

causality), let us also assume tentatively that causal analysis is suitable for at least some purposes in program evaluation and therefore question how qualitative designs might contribute to such analyses.

For the present discussion, it will be helpful to define the quantitative approach to impact analysis as one that relies on the counterfactual definition of causality (a more defensible version of this species of definition is called "factual causation" and will be elaborated below). The counterfactual definition states that X was a cause of Y if and only if X and Y both occurred and, in the circumstances, if X had not occurred, then neither would Y. It is the final "if" clause ("if X had not occurred...") that we call "the counterfactual." We will see momentarily that this conceptualization leads naturally to what we usually think of as quantitative research designs. It is also the source of the well known principle (whose validity will nevertheless be disputed, below) that comparison is necessary to establish causality, the comparison being between evidence for the "result" on the one hand, that is, the occurrence of Y after X, and evidence for the counterfactual claim on the other. Classically, we get our result from observing a group to which a treatment has been applied and our estimate of the counterfactual from a similar group (or a "before" measure) to which the treatment has not been applied. Putting these two sets of observations together so that we have variance on the proposed causal or treatment variable makes the whole amenable to statistical analysis. Design and analysis together yield the classic quantitative package. Note that without data on the counterfactual, a statistical technique such as regression or correlation analysis would not be applicable, even if there were thousands of cases, because there would be no variance on the variable denoting whether or to what extent the treatment was received.

Accordingly, we may provide a specialized definition of the term "qualitative," as well as "quantitative," as they both apply to the idea of design (as opposed, for example, to measurement). A design whose purpose is to determine impact will be considered *qualitative* if it relies on something other than evidence for the counterfactual to make a causal inference. It is qualitative first in the positive sense that it rests on demonstrating a quality—in the present treatment the quality will be a physical connection in the natural world between the proposed cause and effect—and second in the negative sense that it is not quantitative, that is, it does not rely on a treatment *variable*, or on comparing what is with an estimate of what would otherwise have been. A vast number of observations may be and usually are made in a qualitative evaluation, but when it comes to the core of causal inference there is only one *kind* of observation in the perspective of the counterfactual definition, that showing the co-occurence of X and Y. Conversely, even if a study has only two empirical observations—one showing that the proposed cause and effect occurred together and the other showing that when the cause did not occur, then neither did the effect—it is clearly based on *variables* and counterfactual reasoning and will therefore be considered quantitative.

These two approaches would seem to present an interesting and viable picture except for one matter. The qualitative approach does not fulfill the requirements of the counterfactual definition since it does not deal in variables and does not incorporate evidence on the no-treatment condition, but, apparently, no other rigorous definition of causality is accepted in social science (King et al., 1994). If it is indeed true that there is no other, then it would simply be impossible to infer causation based on a qualitative design. Criticisms of the case study as being based on a sample of only one case make exactly this point. On the other hand, there has been at least one approach to causality in the literature on program evaluation that seems on the surface to be both legitimate and to have nothing to do with counterfactuals or quantitative analysis. It may be found in Scriven (1976), where it is called the "modus operandi" method

of demonstrating causality. An explicit alternative definition is not provided there, but the method proceeds as follows:

Y has occurred and the task is to demonstrate that the treatment T caused Y. This is done in part by the process of elimination. That is, there are several other possible causes of Y, such as U, V, and W, and these must be eliminated as contenders in order to settle on T. The process also depends critically on the idea that each possible cause has a known "signature," and if the signatures of U, V, and W can be shown to have been absent while that of T was present, then it may be concluded that T was the cause. The signature may contain events of two types: (a) a known causal chain of events by which T or U, etc., would lead to Y (George & McKeown, 1985); and (b) other events in addition to Y that are known to be associated with an active T, or U, and so on (Campbell, 1979). The elimination of other possible causes is of course a prominent tool in almost all causal inference but it is not enough by itself except in the rare circumstance, such as in randomized experiments, when *all* other possible causes can be identified or controlled. Even then, a reservation remains, as will be shown in a later section. In short, we will see below that while the method of elimination can be extremely helpful in causal inference from qualitative designs, we must rely more on demonstrating that the signature of *T itself* was present and that this proves the causal point. For the present, it is important to see, as Scriven (1976) emphasizes, that the modus operandi scheme is extremely common in everyday causal inference. It is not an arcane, academic conceptualization, but a method of determining causality that is in constant use in the professions and in everyday life, such as in diagnosing what is wrong with a car, or in medical diagnosis, detective work, cause-of-death determination, and so forth. The auto mechanic can infer that a car is overheating because a thermostat is stuck by feeling the temperature of certain hoses. The physician can determine that a certain disease is at work by establishing the presence of two or three critical symptoms.

It is also important to emphasize that this common method is not at all the same as determining causality by counterfactuals. Suppose that a woman has had a heart attack and has died. Probably she would not have died at that time if she had not had the heart attack, but that is not certain. In fact, suppose we could demonstrate that she would have died then anyway from some other cause. Using the counterfactual definition, we would be forced to conclude that the heart attack was *not* the cause of her death because we could not claim the counterfactual (if no heart attack, then no death). But by examining the tissues and otherwise identifying the signature of a heart attack, we might be able to establish the truth, namely, that she did indeed die of a heart attack. Note that in making this determination we do not proceed by bringing forth evidence for the counterfactual—what would have happened had she not had a heart attack—which would in fact lead us completely astray in this case because she would have died anyway. We do not make a dichotomous *variable* out of heart attack, the supposed cause. We consider this to be a one-case study and determine causality by exposing the incontrovertible, physical signature of the heart attack. Finally, it may well be that the signature is known because of prior medical research and also that the prior research in question might possibly have employed counterfactuals, but that does not mean that our causal inference rests on the counterfactual definition. We are not doing that other research at present. We are researching the death of this woman, just as we might research the outcome of a program being evaluated, and we do not use counterfactuals to reach the particular causal conclusion in the specific instance of causality that links this heart attack with this death. To drive the point home, suppose that many impact studies have given us the knowledge that employment can definitely decrease criminal recidivism, just as heart attacks can cause death. If we now do a new study of recidivism and wish to explore a possible causal link with employment, we

have two distinct choices of approach—the quantitative and the qualitative. It does not matter how the prior research was conducted. Both research options are real and independent in the present.

Other research in social science and program evaluation has shown a role for qualitative research in impact analysis, particularly Huberman and Miles (1985) and King et al. (1994). When pressed, however, the core determination of causality in these treatments depends ultimately on the counterfactual method. The best clue we have that a core determination may indeed be qualitative lies in the concept of the signature as it emerges under the various labels of "modus operandi" (Scriven, 1976), "process tracing" (George & McKeown, 1985), and "pattern matching" (Campbell, 1979). The possibility of a qualitative determination of causality, then, has been very strongly suggested and what remains to consider is how this general approach aspires to establish instances of causation without variance on a causal variable (Maxwell, 1996). A totally different definition of causality must necessarily be involved.

The working out of that definition, which is too lengthy to reproduce here in full, may be found in Mohr, 1996. There it is argued that it is appropriate in social science to employ a dual conceptualization of causality. Furthermore, there is abundant evidence that such a dual conceptualization actually is employed. The two facets of the concept are labeled *physical* and *factual* causality, respectively. The first is a relation between events in the natural world, as in "The car skidded sideways and toppled the lamp post on the corner." The second is a relation between statements, truths, or facts (Strawson, 1985), as in, "The fact that the ex-prisoner had steady employment caused the fact that he did not commit further crimes."

Of the two, *physical causation* is more important in terms of the present inquiry, although it is not necessarily the most commonly encountered in program evaluation and social science. Its importance lies in the fact that (a) it is fundamental within the dual approach (it is itself part of the definition of factual causation); and (b) it is the key to causal inference in qualitative research. Physical causality is the common, mechanical or Newtonian sense of causation that is familiar to all. Cause and effect are the relation between a force and a motion, respectively. It is the causation we recognize in the lifting, smashing, pulling, pushing, or tearing of something. The classic example in the Humean tradition is billiard balls on a table, but Hume (1955) sought to exclude this idea as a definition. It is argued elsewhere (Strawson, 1985; Mohr, 1996) that Hume put too much emphasis on the sense of vision and not enough on the sense of touch in reaching this conclusion, but it would be a lengthy and unwarranted digression to review the arguments here. We must proceed as though physical causation were not only intuitively appealing but philosophically sound and refer the interested reader to the cited works for a fuller treatment.

The counterfactual definition of causality, serviceable as it is for some purposes, is well known to be plagued with serious philosophical problems (Mackie, 1980; Mohr, 1996). *Factual causation* is a variant achieved through the interjection of physical causality that preserves the counterfactual definition in spirit while overcoming those problems. The definition is as follows: X was a factual cause of Y if and only if X and Y both occurred and X occupied a necessary slot in the physical causal scenario pertinent to Y. The interjection of physical cause is clear. The spirit of the counterfactual definition is preserved in the idea of a *necessary* slot: if *something* did not occupy that particular functional slot (and it happened in this case to be X), Y would not have occurred. For our purposes, we may treat factual causation as though it were the simpler, counterfactual definition without undue or misleading distortion. The two facets, then, though related, are quite distinct. Physical causation rests on the idea of a direct physical connection in the real world. Factual causation rests on a counterfactual claim and

has no need to refer to a direct physical relation. (For example, in "Because the ex-prisoner was employed, he committed no further crimes," there is no direct physical connection referenced between events in the natural world.)

REASONS AS CAUSES

We have seen that causation can be demonstrated through qualitative research when physical rather than factual causality is the guiding conceptualization. That is why we needed no variables or other quantitative paraphernalia to make an inference about the heart attack, the stuck thermostat, or the diagnosis of the disease. There would seem to be a problem in terms of our central purpose, however, in that the effects of interest in program evaluation and social science are so frequently human behaviors. This is overwhelmingly true in the case of program evaluation, in fact, where common outcomes of interest are behaviors such as self care in nursing homes, criminal recidivism, juvenile delinquency, teen pregnancy, preventive health behaviors, welfare-to-work, substance abuse, overcrowded housing, divorce, safe driving, and safe sex. What are the physical causes of intentional human behaviors? They are at any rate not so tangible or clearcut as hot and cold hoses, or sore throats and fevers. Many if not most who ponder this issue seem to feel that mental states or acts such as decisions, choices, and reasons—especially the latter—are the causes of behavior (Davidson, 1980). We will be guided here by the finding, based in the physiological rather than the philosophical literature, that reasons, in a special sense of the term, are indeed not only the causes but the physical causes of intentional human behaviors (see Mohr, 1996, and works cited there).

Before elaborating the special sense of reasons referred to, let us review the more general sense. Most definitions are similar and suggest two components: If you (1) want to accomplish Y; and (2) believe that action X will do it, then you have a reason for undertaking action X. For example, if you want to improve your agency's productivity and believe that privatizing some of the tasks will do it, then you have a reason for privatizing. So far, the general sense of the concept of a reason does not seem very physical, since wanting and believing are mental constructs. In addition, it is hard to see how reasons so construed can be the causes of behavior. We all have a great many reasons for doing a variety of things at a given time but we do not act on most of them at that time, and some of them are never implemented at all. One might have a good reason, for example, to go to the market for eggs this evening and also to replace a washer in a leaky faucet, but one might carry out only one of those activities, or none. Moreover, a student may see, and even keenly feel that getting good grades is a reason for studying long and hard, but never study very long or hard at all. What does it mean, then, to say that reasons are causes when so many of the reasons that were definitely "there" at a given time did not cause anything?

To answer, it is convenient to introduce the notion of "operative reason," i.e., the reason that actually operated to produce the behavior performed, and to stipulate that the operative reason was different from all of its contemporary reasons in that it was the strongest. If the person fixed the faucet, for example, instead of going to the market, it is because he wanted to stop the drip more than he wanted the eggs. If the student studies little (given a belief in the connection between studying and good grades), it is because she wants something else more, perhaps being accepted as one of an "in" crowd with whom she regularly associates. And if the Director is more interested in pleasing the public employees union than in productivity, the tasks in question may not be privatized this year, or perhaps ever. In this way, in short, one

of the many reasons that may be active at a given time becomes an operative reason and the associated behavior ensues. As a footnote, we see that if a program aims at getting people to behave in certain ways—showing up for work regularly and on time, for example—the program must not only motivate them but motivate them strongly to carry this action out, and if the evaluator wishes to relate the presence or absence of a successful result to the program, it will be important to adduce evidence regarding the strength or weakness of the critical motivation.

The most important aspect of the special sense of "reason" referred to is that, as used here, a reason is an unaware physiological entity and not a mental construct. We all know that we frequently do things intentionally, but without being aware of any particular reason. On the other hand, we are often very much aware of particular reasons for doing things (which we may or may not do). These conscious reasons or thoughts, it is argued, do not cause behavior, although they may at times closely parallel the physiological elements that do. It is well known that for an experience to be stored in and recallable from long-term memory, it must be accompanied by a physical affect residue, a component that records the related affect, if any, at the time of the experience. The "want" part of what we think of as a reason, although we now speak of a "want" that is totally below the level of consciousness, is the positive or negative affect residue attached to some object in memory. The strength of the reason is the electrical or chemical strength of this affect residue. The "belief" part of a reason is the memory of having done some X to achieve Y or something similar in the past, or even of having heard of accomplishing something like Y by doing X. The X is selected and activated when the affect on the relevant Y becomes dominant by virtue of its strength. We see therefore, that operative reasons as rendered here are the *physical* causes of undertaking action X. Finally, it is important for the research implications to note that operative reasons, being unaware, are not necessarily obvious, either to the subject or the investigator. They may frequently be accompanied by a parallel reason or decision that is an aware, mental state, but they also may not, and even if they are, it is the physiological "reason" and not the mental one that is causal. Indeed, when we think sometimes that a certain conscious reason for our behavior was operating we may well be wrong, as we might find out, perhaps, only by careful introspection or with the help of a psychotherapist.

In sum, qualitative research to determine the causes of intentional human behaviors, such as whether or not those behaviors were induced by a program being evaluated, must involve a search for the operative reasons behind the behaviors. In most cases this would undoubtedly involve obtaining information from the subjects whose behavior is at issue. Given that operative reasons are unaware, however, that is not necessarily all that one must do. It might be necessary to obtain information from many other people, from documents, and from the histories of relevant events. These are methods, however, that are familiar in social science, especially in such areas as history, anthropology, and area studies.

CAUSAL REASONING

We have seen that there are two types or conceptualizations of causation, the factual and the physical. When research proceeds as sketched out just above, the causal inference will be based on an argument that makes a case for the operative reason and connects it with the behavior, perhaps with some reservations. The kind of argument involved in that case, as well as that involved in previous qualitative examples such as establishing the connection between

the temperatures of hoses and the overheating of the car, will be called "physical causal reasoning." When the causal inference is to be an inference of factual causation, a different sort of research process is dictated, one that results in the ability to make the argument, again perhaps with reservations, that factual causation has been established. The form of that particular argument would be called "factual causal reasoning."

Factual causal reasoning must necessarily rely on a counterfactual claim. A good deal of energy and design ingenuity would be devoted to being able to argue persuasively that if X had not occurred, neither would Y. This might actually be done within the confines of a single case study, or when talking about a single event that is a minor part of a larger story. That is, one might show that X and Y occurred and argue as best one can on the basis of knowledge of past events, plus a careful and detailed examination of the context and circumstances surrounding the present event, that if X had not occurred here, neither would Y. One might not, that is, purposely make one or more empirical observations of a "control group"—a similar subject and context in which X did not occur. Still, this would be factual causal reasoning and, as the terms are used here, a quantitative design. Or, as featured prominently in the book by King et al. (1994), one might study in depth one or two cases in which X was present and one or two in which it was not. The authors tend to refer to such "small n" studies as qualitative because of the depth of exploration and the manner of data collection and measurement, but in design perspective, these are again quantitative. In most cases, however, factual causal reasoning is based on designs and sample sizes for which statistical methods of analysis become suitable. In that kind of case, it is important to remember from the counterfactual definition that the control group or comparison group or "before" measure, for example, is not of interest in its own right. It is not a matter of comparing what happens (present tense) when X does and does not occur. Rather, the control or similar group is of interest only as a basis of estimating what would have happened in the experimental group if X had not occurred *there*. That is the essence of factual causal reasoning.

Take the evaluation of health education efforts, for example. There we have statistical designs of various sorts in which investigators have concluded that *knowing about* certain risk factors leads to (i.e., causes) certain health related behavior, such as nutritionally sound eating practices (Edwards, 1985). Frequently, there is no thought that this knowledge about risk factors might be a physical cause. Being exposed to information about proper nutrition, for example, is not thought to have *made* the subjects eat properly, but rather we can see that adequately *motivated* people (here is where the physical causation comes in) are much more likely to eat properly if they indeed know what is proper than if they do not—accurate knowledge is more auspicious than ignorance or misconception. In any case, the reasoning proceeds by showing that those who received and absorbed the health education materials tended to eat properly. Those who did not, however, or who did not understand it or remember it, tended not to eat properly, thereby indicating that the knowledge in the first group was a necessary condition. (Mohr, 1995, p. 267)

Physical causal reasoning, on the other hand, is focused on demonstrating persuasively that a particular kind of physical connection took place between two events in the real world, in essence, the connection between a force and a motion. There is some basis for considering that it is this sort of reasoning, in contradistinction to factual causal reasoning, that gets at *real* causes. Frequently, in other words, we are not satisfied with evidence from observational studies or even randomized experiments because the real, that is, the physical causal mechanism remains a mystery. The physical evidence may frequently be required to support the statistical, and it will almost always contribute to a more convincing package if it can be obtained.

One excellent example is the long and persistent interest in the question of how, physiologically, smoking operates to cause lung cancer. We have been bothered, in other words, by the fact that the signature of smoking has remained obscure. In contrast, it took us less time to become convinced of the effect of fluorocarbons on the ozone layer because we not only had statistical time series relating the two, but an account of the chemical processes involved, as well.

To elaborate in the context of more standard program evaluation, consider that we might be trying to learn whether to attribute a lower fatality rate from motor vehicle accidents to a well-publicized police crackdown on speeding in one community. Application of the before-after or the interrupted time series design to the fatality rate is one option. Comparison to a community that is similar and that has had no crackdown is another. Both of these are in the classic, quantitative or factual-causal mode and both are weakened by the standard threats to the validity of quasi-experiments. An alternative to be pursued instead, or perhaps better yet in addition, would lead us to confine our attention to the one community and the one time period of concern—a case study insofar as this particular aspect of the design is concerned. We might use a survey to determine not only whether and to what extent but *how* the citizenry became aware of the existence of the crackdown. That is, we would *not* try to show that they would never have heard of the crackdown if not for the media used by the project (factual causal reasoning), but rather to document that they actually did read about it here, or did see it on TV there, or were told about it by a police officer, or by a friend who claimed to have seen the TV spot, and so forth—all physical processes. Furthermore, we would be interested in understanding the behavior itself—not only *whether* the citizens consciously drove more slowly, but why—was it a heightened sense of the dangers of speeding, for example, or a heightened subjective probability of getting caught? Note that the quantification involved in both of these links in the project, given the survey involved, would be just a matter of counting citizens to get an idea of the prevalence of the causal processes uncovered. The quantification would not in any way be part of a *design* to permit causal inference. The design is unmistakably qualitative. No evidence for the counterfactual would be brought forward because the survey would not contain any. On the contrary, whether one citizen is surveyed or a hundred, the design is critically dependent on persuasively establishing the occurrence of the two physical causal processes in the individuals concerned—learning about the crackdown because of project activities and modifying behavior because of strong motivations sparked by the project's messages.

Another brief example might be helpful in suggesting the kinds of roles available for physical causal reasoning. In the kind of research exemplified by eating practices and knowledge of good nutrition in the quoted paragraph above, the efforts to increase knowledge would necessarily be evaluated by factual causal reasoning alone. The nutritional knowledge simply was not a physical cause. As suggested, however, there would have to be separate motivations, or reasons, impelling the sound dietary behavior apparently advocated. A more thorough program would seek to inculcate these, as well. An impact analysis of the program might then proceed either by factual or by physical causal reasoning or both. There is no reason, in other words, why the two forms of logic or research design cannot both be brought to bear (a) within the same evaluation, as in evaluating the efforts to impart knowledge by factual and motivation by physical causal reasoning; or (b) on the same outcome, as long as we are concerned with a physical cause, as in employing two different designs to determine whether or not people changed their eating habits because of the program's efforts to persuade them of the benefits.

The purpose of this paper is in large part to identify the idea of physical causal reasoning and to suggest its functions. The details of successfully implementing such reasoning in evaluation and other research are beyond the scope of our efforts here and no doubt will only be accumulated over many trial projects. Something about the basic elements of implementation, however, rises out of the definitional logic of physical causation itself. This conceptualization of causality has to do with the relation between a force and a motion. With regard to human behavior, these components are bound up in the operative reason and the behavior itself, respectively, as in the relation between fear of getting caught by the police and keeping within the speed limit during the crackdown. To establish this relation persuasively, four elements are apparently important. One must show first that the person or group studied must have had the reason to which we want to attribute causality—sometimes whether they themselves claim it or not. Second, it is important to make the case that this reason was very strong relative to other considerations that could have been active at the time. Third, it is generally crucial to show by various other manifestations that it was indeed this reason that was operating—tracing out the "signature" of the reason in the style of the modus operandi approach. And last, it will generally be helpful to establish that it is proper to *interpret* the actors' behavior as intentionally keeping below the speed limit, for example, or following sound nutritional practice, and not just as driving at a certain speed or eating certain foods, as would be common in quantitative research.

It would be a digression to discuss at length the various other perspectives on causality and causal reasoning that have been suggested in relation to qualitative research, such as constructivism, neorealism, and so forth. It may be well, however, to emphasize one point, and that is that what is offered here are *definitions* of causation such as must normally support all such perspectives. If one takes a constructivist view of "causation," therefore, or emphasizes "webs of causation" rather than single, linear, cause and effect dyads, or aims with the neorealist at discovering "causal mechanisms" rather than causes, one can always be challenged to specify what one actually means by the root term "causation" or "causal" in one's own particular context. Exactly as was the case with the modus operandi method, these other perspectives do not obviate the need to conceptualize causality, nor do they automatically provide the required conceptual definition in themselves. They tend to use the term "cause" as though it were previously defined. It is suggested here that the response to this challenge to define can always be rendered in terms of one of the two conceptualizations proposed above. It is therefore suggested further that the two types of causal reasoning, physical and factual, will in that sense always be pertinent, whatever the philosophical framework, although they may not always be important enough to an author's point to warrant attention. Finally, although qualitative research is firmly established as a scholarly approach in social science and program evaluation, still, consideration of the notion of causality seems to be ruled by the counterfactual definition, which, from the standpoint of design, is basically quantitative in orientation. Thus, the possibility of thinking in physical causal terms should open up notably wider prospects for thought, particularly within the qualitative regime.

INTERNAL AND EXTERNAL VALIDITY IN THE QUALITATIVE APPROACH

Let us first accept the term *internal validity* to mean the validity of a causal inference based on the research, as was the case when the term was first used by Campbell and Stanley (1966). In this light, the concept has not been very relevant for qualitative evaluation in the past

because qualitative evaluation has not been much preoccupied with causality. Given that the present paper has been vitally concerned with the marriage of "qualitative" and "causal," however, internal validity does become germane as an issue.

There are two components to the practice of establishing internal validity in quantitative research, and the same two will hold for qualitative research, as well. The first has to do with showing the fulfillment of the governing idea of causation through the *design* and the operations of the research. In the quantitative mode, this primarily means arranging to produce evidence for the counterfactual (since establishing that X and Y both occurred is generally not much of a challenge)—as for example by the device of a comparison group. In the qualitative mode, it means arranging to produce evidence for a certain sort of physical relation between X and Y. In both cases, however, the evidence will never be unassailable, so another important component of establishing internal validity becomes the attempt to rule out certain other, plausible causes. Even in the randomized experiment, the evidence for the counterfactual is only probabilistic, and not certain, so it is common to achieve greater certainty on at least some dimensions by introducing control variables. In both approaches, therefore, ruling out alternative causes occupies a prominent place in the quest for internal validity, but the deductive logic of the process of elimination does not obviate the role of the factual causal argument in the quantitative approach, nor does it eliminate the role of the physical causal argument in the qualitative approach.

In this light, we see that the elimination of other causes in the modus operandi method is not meant to be the same as establishing evidence for the counterfactual, which would be mixing the two forms of reasoning, and indeed it is not the same. First of all, one could in no way claim to prove the counterfactual by that method unless one eliminated *all* other causes, and it will always be impossible to make that claim legitimately. Second, even if we knew that nothing else caused Y, we would not necessarily know that nothing else *would have* caused Y. This is illustrated by our prior example of the woman and the heart attack. Recall that we assumed for the sake of argument that she would have died then anyway, even if she had not had the heart attack. Suppose that the woman had been found dead in her car, which had crashed into a tree. The question is whether she had a heart attack and died, crashing into the tree while the death was in process, or if she died of a concussion as a result of crashing into the tree. We might be able to tell by careful autopsy that she died of a heart attack, so that all other causes are categorically eliminated, but it may in fact be true that if she had not actually died of the heart attack, she would surely have died of the concussion or other injuries. To eliminate all other possible causes, therefore, as by finding the incontrovertible signature of the heart attack, does not establish the truth of the counterfactual—here, she actually would have died anyway. In quantitative research, the process of elimination is meant rather to strengthen a basic claim for the counterfactual that will generally have weaknesses. In qualitative research, it is meant to strengthen the basic claim of a certain physical connection—for example, in striving to rule out excess greenhouse gases resulting from forces independent of industrial and other human activities. In fact, it will by no means always be necessary to resort to the effort to identify and eliminate other plausible causes. The design-based claim for the counterfactual or the physical relation may suffice. Just as in a randomized experiment, and even in some quasi-experiments, we might frequently consider the case for the counterfactual to be strong without bothering to try to identify other plausible causes and enter them as control variables, so in a qualitative study we might consider the case for the physical relation to be strong, or obvious, without bothering to try to identify other plausible causes and show the absence of their signatures. For example, if we show that certain drivers became aware of the

police crackdown because they got tickets and the officers told them about it, we do not necessarily have to speculate on how those particular drivers might have become aware of the crackdown otherwise, but did not.

Nevertheless, many will say that internal validity in the qualitative approach must be inherently weak. Usually, the claim that a certain reason was the operative reason for a behavior, for example, must in the end be speculative. In that approach, there is nothing like the laws of probability to give credence to an inference. This reservation is valid and important, but it needs to be put in perspective. The difficulty is relative. Other approaches have similar problems, and weak internal validity for a method has never been a good reason to reject it categorically. In particular, one cannot make an airtight case on the basis of factual causal reasoning, either. There is always a necessary residue of subjective assessment. In randomized experiments, the threat of contamination (impure treatment delivery or inadequate control of the environment) can never be extinguished altogether, nor can the reservation that the statistical causal inference carries a probability of error equal to the significance level. At the 5% level, we *will* make errors about 5% of the time. In quasi-experimental designs, which are extremely well represented in evaluation journals, the threat of selection bias is similarly persistent and is in fact notoriously troublesome. Furthermore, ex-post-facto or observational studies are well known to be pathetic in terms of internal validity, yet they are in constant use, probably because there is often much to be learned from them nevertheless. On the other hand, just as we have imperfect but substantial confidence at times in the conclusions of experiments, quasi-experiments, and observational studies, we may often have substantial confidence in the case made for a physical cause through the modus operandi method or other physical causal reasoning, whether the subject be the overheating of an engine, the penetration of a message by means of the mass media or a policeman, the controlling desire to avoid speeding tickets or, let us say, the controlling desire to avoid the inevitable future hassles with bureaucrats by immediately correcting violations noted on a safety inspection. Reasons and other physical causes that are not obvious may nevertheless frequently be demonstrated with considerable satisfaction and edification.

This is not to say that further constructive thought regarding internal validity is either futile or unnecessary in the context of qualitative designs for impact analysis. On the contrary, much should be done and there is in fact evidence that the issues are not intractable. For example, as will be reviewed momentarily, Lincoln and Guba (1986) have summarized much that has been done to conceptualize for qualitative research what would seem to be analogous to issues of validity in quantitative research. Much of the collective conceptualization has more to do with qualitative approaches to measurement or description rather than design, but some concepts, credibility for example, are clearly applicable to the present discussion and make a stimulating contribution. The whole idea is both new enough and challenging enough, however, that we should not expect definitive analyses to have been produced already. Finally, Mohr (1999) has suggested the criterion of a "consummate causal understanding" that should also be helpful in developing methods of achieving and testing for internal validity in the qualitative approach to impact analysis. A consummate causal understanding of the past behavior investigated is achieved when all pertinent challenges and questions regarding that past behavior and one's explanation of it are answerable from the data presented and their logical extensions.

The issue of *external validity* must also be considered, but the first observation to be made, it would seem, is that it is important to keep the two types of validity distinct. Because a sample size of just one case appears to threaten both types, there may be a tendency to

blend them together when critiquing the case study as a design. "Let us accept," the critic might suggest, "that causality can be established with a sample of one, but is it not better to have a large sample? Is it not better to observe many causal instances rather than just one or two? The one or two might have been unique." Perhaps so. We proceed now to consider exactly that issue. But it is necessary to recognize at the outset that it is an issue of external validity and not internal. If causation is established then causation is established. The question whether it is better to establish it in more subjects rather than fewer is purely an issue of generalizability.

It is well known that external validity is critical (Cronbach 1982). There is hardly a point in finding out for sure that T caused Y if we have no interest in applying the finding either to the same subjects in a subsequent time period or to other subjects. The strongest case for internal validity, however, guarantees nothing in regard to external validity. True, if the observed subjects were randomly sampled from a larger population we can generalize the results to that population, but (a) that pertains to internally valid physical as well as factual causal results; and (b) such a generalization is still limited to that particular larger population and, even more restrictive, to the past time period in which the subjects were observed. In particular, there is no basis for claiming that a result is theoretically valuable just because one can generalize it to a certain population that has been randomly sampled.

Since internal validity is in no sense an aid to establishing external validity, it follows that factual causal reasoning is no more auspicious for external validity than physical causal reasoning. Both are oriented toward causality, not generalizability. As noted previously, for example, if we used old or impressionistic evidence to reason for one individual subject that if *T* had not occurred, then neither would *Y*, the mode would definitely be one of factual causal reasoning (quantitative reasoning, that is, even in the single case), but it is hard to see that this would give us any particular basis for generalizing the results to other subjects. We see, therefore, that it is not the design that leads us to prefer quantitative research over case studies as a basis for generalization, but the mere fact of the number of subjects involved, whether people, schools, cities, or countries. It will usually be more comfortable to generalize when many subjects have been causally affected by a treatment than when only one or two have been affected. The one or two might be unusual, even unique. Of course, the many subjects observed in a large sample may also be quite different from others to which one would like to generalize, and in fact they generally are, as is evidenced by the fact that one can almost always find interacting variables of interest that qualify a relationship. There is therefore some advantage in numbers, but it should not be oversold. Conditions change over time. One subculture is different from another. One school is even different from another within the same district, so that one may very well not observe the same regression coefficients as one replicates the analysis.

At the same time, qualitative studies have their own advantage with respect to external validity that has long been recognized. It is that the study in depth should enable one to *understand the process*, and there may be no factor as important as this for generalizability in social science. The more thoroughly we understand the causal process by which a treatment has affected an outcome in one case, the better the position we will be in to know when a similar outcome will result in another case (cf. "extrapolation" in Cronbach, 1982; see also Lincoln & Guba, 1986). Thus, the quantitative approach generally has an advantage in numbers and the qualitative approach generally has an advantage in understanding. It is unfortunately difficult to capture both simultaneously because they tend to work against each other in practical terms. There is no reason why quantitative studies cannot

strive for depth, but limited resources will constrain the effort. Similarly, physical causal reasoning can be carried out on a large sample, as the example of the survey in connection with the crackdown on speeding was meant in part to show, but again resources constrain depth when numbers are large. We can conclude regarding external validity that it is well to be sensitive to both advantages—numbers and understanding—regardless of the approach one takes, but the primary point to be made in the present context, perhaps, is that by no means is the qualitative approach to impact analysis inherently inferior in terms of external validity.

In this section, physical causal reasoning has been scrutinized in terms of internal and external validity. One might well ask, in addition, how it would fare on criteria of appraisal that have been more specifically devised for qualitative evaluation research. This area does not have as long a history as internal and external validity and is not as well developed. We may briefly consider, however, the overview by Lincoln and Guba (1986) of some of the major suggestions and ideas. They are summarized into the two umbrella categories of "trustworthiness" and "authenticity."

It appears that some of these ideas, at least, are applicable to physical causal reasoning and that the latter can be expected to score high on such criteria in a well planned and executed study. Under trustworthiness, in particular, the category of transferability is suggested as a parallel criterion to external validity. As indicated above in the discussion of "understanding the process," the criterion of the capability of the transfer of results from an observed application to a potential one is a strong form of external validity and, perhaps, the strongest form that can be achieved in social science, where universal laws are dubious. Similarly, the whole idea of "audits" that has been developed in connection with qualitative research is fully applicable to physical causal reasoning and would seem to be an effective way to insure against any unavowed and unwelcome intrusions of personal values and biases. Finally, when applied to qualitative design, Lincoln and Guba's category of "credibility" would seem to have almost the same meaning as internal validity in more traditional contexts. The issues that are important to consider in this regard were summarized earlier in this section under the rubric of internal validity.

In the category of "authenticity," Lincoln and Guba (1986) discuss a number of criteria that are not meant to be parallel to the criteria traditionally applied to quantitative evaluation but grow rather out of the specific characteristics of qualitative evaluation. These criteria, such as fairness, educative authentication, and catalytic authentication are less applicable to the approach outlined here. The reason is that they are largely bound up with alternative goals for evaluation, beyond impact analysis. In some dimensions they should be important for all evaluation, including quantitative, as for example the importance of obtaining fully informed consent from human subjects and informants or of considering the values of all stakeholders rather than just a few when assessing the merit or worth of a program. For the most part, however, they are not only pertinent primarily to goals and functions of evaluation other than impact analysis, but at times they articulate those very goals. For example, according to Lincoln and Guba, "It is...essential that they come to appreciate...the constructions that are made by others and to understand how those constructions are rooted in the different value systems of those others" (p. 81). Authenticity, in sum, seems less applicable to the physical causal approach in qualitative evaluation than trustworthiness, but that is not surprising when the central role of causation is considered.

CONCLUSION

Physical causation as a way of explaining the world seems to come naturally. There is ample room in the practice of program evaluation for using it to demonstrate the causal nature of one or more links on an outcome line. In particular, in-depth study and interpretation are well suited to discovering the true or operative reason behind a behavior. In this introductory treatment, the purpose has not been to show how qualitative evaluation is done, but rather to show how awareness of an alternative causal epistemology may be coupled with current knowledge of qualitative methods of data collection and measurement to fashion a sound, independent approach to the analysis of impacts.

ACKNOWLEDGMENTS

The text of this paper is based in part on Mohr (1995), Chapter 11. A prior version of the paper itself was delivered at the annual meeting of the American Evaluation Association, Vancouver, November 1-5, 1995. I wish to express my appreciation to the editor, to his editorial assistant, to Jennifer Greene, and to three anonymous reviewers for helpful comments and criticisms on an earlier draft—to which, no doubt, I have not done full justice.

REFERENCES

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Chen, H. T. (1990). *Theory-driven evaluations.* Newbury Park, CA: Sage.

Cook, T. D., & Reichardt, C. S. (Eds.). *Qualitative and quantitative methods in evaluation research* (pp. 49-67). Beverly Hills, CA: Sage.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco, CA: Jossey-Bass.

Davidson, D. (1980). *Essays on actions and events.* Oxford: Clarendon Press.

Edwards, P. K., Acock, A. C., & Johnston, R. L. (1985). Nutrition behavior change: Outcomes of an educational approach. *Evaluation Review 9*(4), 441–460.

George, A. L., & McKeown, T. J. (1985). Case studies and theories of organizational decision making. In L. S. Sproull & P. D. Larkey (Eds.), *Advances in information processing in organizations.* R. F. Coulam & R. A. Smith (Eds.), *Research on public organizations.* Vol. 2. (pp. 21-58). Greenwich, CT: JAI Press.

Goddard, A., & Powell, J. (1994). Using naturalistic and economic evaluation to assist service planning. *Evaluation Review, 18*(4), 472–492.

Greene, J. C. (1994). Qualitative program evaluation: Practice and promise. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 530-544). Thousand Oaks, CA: Sage.

Greene, J. C., & Caracelli, V. J. (Eds.). (1997). *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms.* New Directions for Program Evaluation, No. 74. San Francisco, CA: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches.* San Francisco, CA: Jossey-Bass.

Huberman, A. M., & Miles, M. B. (1985). Assessing local causality in qualitative research. In D. N. Berg & K. K. Smith (Eds.), *The self in social inquiry* (pp. 351-381). Newbury Park, CA: Sage.

Hume, D. (1955). *An inquiry concerning human understanding*. C. W. Hendell (Ed.). New York, NY: Liberal Arts Press.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Lincoln, Y. S., & Guba, E. G. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. In D. D. Williams (Ed.), *Naturalistic evaluation* (pp. 73-84). New Directions for Program Evaluation, No. 30. San Francisco, CA: Jossey-Bass.

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

Maxwell, J.A. (1996). Using qualitative research to develop causal explanations. *Working Papers*. Cambridge, MA: Harvard project on schooling and children.

Mohr, L. B. (1995). *Impact analysis for program evaluation* (second ed.). Newbury Park, CA: Sage.

Mohr, L. B. (1996). *The causes of human behavior: Implications for theory and method in the social sciences*. Ann Arbor, MI: University of Michigan Press.

Mohr, L. B. (1999). One hundred theories of organizational change: The good, the bad, and the ugly. In H. G. Frederickson & J. M. Johnston (Eds.), *Public administration in a time of turbulence: The management of reform, reinvention, and innovation*. Birmingham, AL: University of Alabama Press, forthcoming.

Reichardt, C. S., & Cook, T. D. (1979). Beyond qualitative *versus* quantitative methods. In T. D. Cook & C. S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 7-32). Beverly Hills, CA: Sage.

Scriven, M. (1976). Maximizing the power of causal investigations: The modus operandi method. In G. V. Glass (Ed.), *Evaluation studies review annual*, Vol. 1 (pp. 101-118). Beverly Hills, CA: Sage Publications.

Shadish, W. R., Jr., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Stake, R. E. (1981). Case study methodology: An epistemological advocacy. In W. Welch. (Ed.), *Case study methodology in educational evaluation*. Minneapolis, MN: Minnesota Research and Evaluation Center.

Strawson, P. F. (1985). Causation and explanation. In B. Vermazen & M. B. Hintikka (Eds.), *Essays on Davidson: Actions and events* (pp. 155-136). Oxford: Clarendon.