Evaluation in Education

CURRENT APPLICATIONS

Edited by

W. James Popham

University of California,
Los Angeles

McCutchan Publishing Corporation
2526 Grove Street
Berkeley, California 94704

Contents

Figures

1

Evaluation Perspectives and Procedures

Michael Scriven
University of California, Berkeley

MAXIMIZING THE POWER OF CAUSAL INVESTIGATIONS: THE MODUS OPERANDI METHOD

Granted that control-group studies (including possibly self-controls) are the method of choice for studies of social intervention (Campbell, Tatsuoka, Mosteller, Meehl), we must frequently face the need to do the best we can with nonexperimental data. I want to present a sketch of what is sometimes the best methodology for this, reconstructed from the procedures of the historian, the detective, the anthropologist, and the engineering troubleshooter. This methodology has been completely neglected in evaluation theory, though not completely ignored in practice since anthropologists and historians have occasionally been used as evaluators. I believe that a reasonable account of these methods would make them accessible to the more traditionally trained and oriented evaluator, especially as a substitute for the usually more desirable experimental and quasi-experimental approaches when these approaches are impossible. And, in almost every investigation, they would serve as a supplementary device to increase the reliability of conclusions reached by the more common designs.

It is not adequate to increase the use of this approach by using more historians and anthropologists (or clinicians) in evaluation teams, although this is not a bad idea. What is more important is that the method itself become an explicitly formulated and understood tool in the repertoire of every investigator, for in most evaluations there are no resources for auxiliary personnel. Moreover, scholars whose field requires them to depend on the modus operandi (MO) approach often cannot articulate it well and, hence, do not interact too well with scholars to whom the approach is strange. (For general background, see the discussions in P. L. Gardiner, ed., *Theories of History* (New York: Free Press) and in William Dray, ed., *Philosophical Analysis and History* (New York: Harper and Row).

The procedures outlined here involve more than a formalization of MO analysis; they represent a set of causal inference patterns of which MO analysis is probably the most distinctive. Full conversion of the modus operandi method into quantitative techniques may or may not be possible, but some suggestions as to procedure follow. Some conversion may already have been done in as-yet-unpublished developments of path analysis, from the Wisconsin conference or elsewhere. But I have been so struck by the excessive crudity and impracticality of early work in path analysis, and by the continued neglect of MO methodology by sophisticated investigators, that I felt it would be worth spending some time trying to do the job in my own way. I would be most grateful for criticism including references to earlier methodological discussions of this approach. It is, of course, related to John Platt's "strong inference" and to Gilbert Harman's "inference to the best explanation," though it antedates both.

An earlier version of this section was commissioned by, and presented orally at, the Battelle Institute in Seattle, Washington.

Statement of the Problem

Assume the general form of the problem is to identify the cause or causes of phenomenon X. Of particular interest is factor A, which represents an earlier intervention that it is our task to evaluate. Factor A will be deemed successful (or unsuccessful) to some degree if A caused X (and certain other conditions are met), not otherwise. X is thus some effect which is demonstrably meritorious (or undesirable) such as learning gains (or increased vandalism). X may of course be a change in the value of a variable, or, atypically, the absence of such a change; or the appearance of a new variable or configuration of variables. It should be noted that the identification of causes is of crucial importance not only for intervention studies (planning social change), but for "pure understanding," which is often the historian's concern. We are thus searching for the cause of X (general problem) or testing the hypothesis that A was the cause (special case).

Now cause hunting, like lion hunting, is only likely to be successful if we have a considerable amount of relevant background knowledge. The most general proposition in this stock of background knowledge, the most general presupposition of our search, must be a claim of weak local determinism, which means that we must have grounds for supposing that X's (or X's in these circumstances, C) are usually caused. Such an assumption fails for many phenomena in the realm of elementary particles. It is well supported in much of the social domain, but it fails even there with respect to certain decision behaviors (by analogy with the argument in my "An Essential Unpredictability in Human Behavior," in *Scientific Psychology*, ed. Benjamin Wolman and Ernest Nagel [New York: Basic Books, 1965]).

Note that we do not require absolute (that is, exceptionless) as opposed to limited or statistical determinism since all we can hope for anyway, given limitations of measurement, is a probable conclusion and we can get that from the statistical ("weak") version of determinism. Methodologically, as long as X is *probably* caused, we have to act as if it *is* caused, on minimax grounds (consider also the optimal betting strategy for a coin known to have slight bias toward a known side). And we only require local determinism (limited to X-like entities, possibly only in specified circumstances, C) as opposed to general or universal determinism (that is, a determinism of all sociological phenomena or all phenomena).

Now most automobile mechanics (and many historians) are not interested in abstract determinism; instead, they need something more specific. They rely, although they may not realize it, on a claim that entails the one just mentioned, namely: Most X's (in C) are caused by A, A′, A″, . . . , A^n. Let us call this a *quasi-exhaustive causal list.* The sense in which they need this does not imply that they can state it, but it is easy to prove that they know it without relying on verbalization (a distinction overlooked by the Educational Testing Service [ETS] in constructing the national auto mechanics test first used this year), and only a little less easy to show that they need it, as we will see. What is interesting about these lists is that they are rather modest claims by contrast with "determinism," that is, strong general determinism, which is usually said to be a required assumption for scientific investigation. As a matter of fact, something can still be done even if we only know some of the entries in such a list.

Causal List Inference

Given the preceding background data, we can sometimes determine the cause of X very simply, by means of the "presence check." We merely check for the presence of each of the A's, and hope that only one is present. The inference is then easy. (See note at bottom of page 84 concerning this inference.)

Almost all A's are caused by A, A′, A″, . . . , A^n.
X has occurred.
A did occur.
A′, A″, . . . , A^n did not occur.

(Probably) A caused X on this occasion.

The probability can be very high if the causal list covers nearly all cases with few causes, and we can often give such lists.

If no A from the list occurred, we strive to discover the cause of this particular X by applying causal lists about analogous phenomena; and, if we do so, we are able to add an extra A to the list for X and increase our confidence in its completeness. If we do not discover the cause of X after such an investigation, and such cases recur, we have to reduce our confidence in the list's completeness. Since the third premise is not necessary, checking it provides some confirmation of the conclusion or increases the probability of the conclusion. This is an important source of internal confirmation and can be cast in classical predictive form if we leave the check for A's presence till the end. The absence of the other possible causes then generates the prediction of A's presence which is then confirmed.

Modus Operandi Inference

If more than one A occurs, we move to the modus operandi check. Since it always provides some further confirmation (probability increment), it is good practice to use this check anyway. Suppose that both A and A′ were present. There are four possibilities of interest: A was the (sole) cause, A′ was the (sole) cause; A and A′ were co-causes; neither was a cause. We now move toward discrimination among these alternatives.

The MO of a particular cause is an associated configuration of events, processes, or properties, usually in time sequence, which can often be described as the *characteristic causal chain* (or certain distinctive features of this chain) connecting the cause with the effect. In its most common usage, in criminalistics, the term refers to such characteristics as the method of entry used by a burglar, the kind of weapon and occasion used by a hit man, the communications procedure used by kidnappers. In the autopsy context (X=death), to go one step further, the MO's for drowning, poisons, and heart failure are well known; the art or science of differential diagnosis for the toxicologist partly consists of MO analysis. It is not the only technique, however, although the line of demarcation is not sharp. If the detective notices the smell of bitter almonds (one of the few experiences which, apparently, everyone can be said to recognize without having any memory of it), he is not using MO analysis; if he notices the characteristic facial rictus, he is. The first is a property of the poison; the second is one of its effects.

In cases where the cause occurred substantially earlier than the effect, MO analysis is easily distinguished from identification of the presence of the cause in that it refers to processes occupying the temporal gap between cause and effect. But we are often not so fortunate as to have such an interval.

The basic truism for MO analysis is that only real, that is, operative, causes fulfill their MO "contracts." Even if A and A′ are both present, which we may determine directly or by inference from certain MO cues, one may not have completed the causal connection to the effect. The victim was poisoned and shot, but it was the shot that killed him because we do not find the knotted muscles that would be associated with death caused by an alkaloid from this group, although we do find that poison in the stomach and some of the early symptoms of its ingestion such as crossed eyes. We will often use "MO" to refer to the whole causal chain, rather than just to the most distinctive features of it.

The general nature of our task is thus one of pattern recognition or, in the language of educational psychology, configural scoring. In general terms, this part of the investigation focuses on discovering how many *complete* MO's are present. Thus, the total sequence of tasks, inferences being of course probabilistic, is as follows:

(i) Check for the presence of each A. If only one, that is the cause.
(ii) If more than one is present, check for complete MO's. If none, then none of those A's was a cause.
(iii) If only one MO is complete, the A with which that MO is associated is the cause.
(iv) If more than one complete MO is present, the associated factors are co-causes.

Pattern-recognition computer programs rely heavily on the fact that they only have to discriminate between a finite set of possibilities, and this makes cause hunting feasible (just as it made possible Mosteller's brilliant solution to the problem of determining authorship of the Federalist Papers, a typical MO approach). Hence, the reliability of the MO methodology depends upon the reliability of the causal list since it provides the candidates. Anyone who has worked from a trouble-shooting chart knows that very high reliability, even without an exhaustive causal list, can be attained in the mechanical-electrical-medical domain. The same is true of criminalistics. What about the educational-clinical-social context? And what about looking for the causes of desirable effects, as opposed to trouble? I would say the situation is closely comparable, although this fact is greatly obscured by the tendency to think that gross differences in predictive power between these groups of disciplines has a bearing on the explanatory situation. The fact remains that causes are, by their nature, explanatory and not predictive factors although we are sometimes lucky enough to get the predictive power as a bonus. It is not hard to list, and it is easier still to recognize, most and often nearly all of the likely causes of a given, substantial, and highly specified social phenomenon, be it recession, depression, regression, or delectation, in particular forms and circumstances.

Even where that is impossible, there is a reserve position. Partial causal lists are still useful, even though they will not support a definite eliminative analysis from mere presence checks, as does the quasi-exhaustive list. Both lists support a highly probable conclusion if the MO check is both complex and successful. The antecedent likelihood that an unknown factor will have the same MO is, in general, very low. Thus:

(i) A and A′ can sometimes cause X.
(ii) Nothing else is known to cause X.
(iii) A but not A′ was present.
(iv) The MO of A, which is highly distinctive, was present.

A probably caused X.

Again, the third premise can be omitted, but, with this weaker inference, its presence is valuable. The sense of "highly distinctive" that is needed for validity here refers to unknown other possible causes of X. Since they are unknown, the safest inferences will be those where the MO configuration is very complex.

A hybrid case, which reflects a good deal of acute implicit reasoning, involves a first premise that lists known possible causes and also hypothetical or speculative possible causes that have never actually been demonstrated as such. The combination might conceivably yield a quasi-exhaustive list, but it is more likely simply to improve the reliability of the conclusion, just as it would if the extra causes were known to be capable of producing X. There is often some problem in deciding what "this phenomenon" is, when trying to give a list of all possible causes of it. After all, there may be something special about this occurrence of "it," which, if known, would preclude the possibility that one of our candidates could have done it. The argument just given, which shows that the inference is still possible even with entirely speculative candidates, entails that there is no need to agonize over such a question. One should simply include any possible cause, in any pragmatic sense of "possible," with either weak theoretical (or analogical) or direct empirical support for its candidacy. The cost of extra length in the causal list is minor, and the gains may be large.

We are thus able to get plausible conclusions from incredibly weak and commonly available premises. Does this have practical applications for the design and support of evaluations? I believe it does. I believe that the main thrust of efforts towards sophistication should now turn from the quasi-experimental toward the modus operandi approach.

Consequences for Research Design and Ideology

One direct consequence of making MO inference patterns explicit is the suggestion that our experimental designs should incorporate what one might call "signature arrangements," "tracers," or "telltales," which will expose distinctive characteristics of the causal candidates of interest. Another is that more attention needs to be given to externalizing the implicit knowledge of causal lists and MO's possessed by many specialists, be they master teachers or union leaders. In fact, these points connect with and mutually reinforce a larger perspective. I think the time has come to change our orientation in the development of social science away from the goal of abstract, quantitative, predictive theories toward specific, qualitative, explanatory checklists and trouble-shooting charts. I think that such a large-scale change is quite closely related to the simple models of argument discussed above, for the ideology of competing methodologies is usually expressible in very simple formal models (think of the Hypothetico-Deductive Model, Covering Law Model, and others). It is not that evaluators, social psychologists, sociologists, and others have never done this kind of fine-structure analysis. Rather, they have done it, rarely and unsystematically, as a work of supererogation or with a degree of informality that leaves it in the category of anecdotal evidence, and it has been done with little or no assistance from the statisticians whereas it should have been a major focus of attention and development.

There are many complications and implications of this account that need careful consideration. Most notably, there are three species of overdetermination that lead to refinements of the MO procedure, and there is the fundamental problem of the origin of the causal lists, their relation to more conventional laws, and the alleged ultimate connections of causation with prediction manipulation and determinism. I have discussed some aspects of these, and other related issues, in "The Logic of Cause" (*Theory and Decision,* II [No. 1, October 1971, 49-66]).

Application to Evaluation Design

Suppose we wish to evaluate the performance of a small-scale bureau we set up on a campus to improve undergraduate and other teaching. We cannot count on the interrupted time-series approach because there are too many novel, powerful, and erratic external influences on the dependent variables at this time, among them, pressure from professional associations and legislatures to improve teaching. We cannot run a proper control group because of moral considerations (we should not withhold help in this area) and contamination. We could try ex post facto matching, but the main problems are the need for a large sample, which is certainly too expensive and may be impossible on a small campus, and the essential bias of an ex post facto study, which is likely to have serious substantive significance.

If the MO approach is used, we first set up a signature or tracer system, trying to arrange things so that the procedures recommended by the bureau are distinguishable from those coming from other sources at least in minor, nonfundamental ways. Where the procedures are novel, this presents no difficulty. If another source has the best available solution, for example, to the problem of designing a student evaluation questionnaire, insignificant changes can be made in the wording on copies distributed by the bureau, and they can be distributed without cost to facilitate getting the tracer into the bloodstream. We try to ensure, and here the intrusiveness may be a small trade-off cost, that verbal counseling is accompanied by written materials with tracers in them or that the verbalized content itself uses a novel, preferably useful, term or two. Monitoring the bloodstream of information through the university later, we can detect the passage of "signed" material and assess deterioriation, implementation, and so forth.

Focusing on MO in another way, we identify a sample of information couriers or influence peddlers, people using bureau services and exhibiting some evangelical zeal; they are interviewed on a regular schedule to catch diffusion effects, and the hypotheses generated by the interviews are followed up. They (or others) may be encouraged to keep logs of dissemination transactions. (Here we find a useful reinforcement between evaluation procedures and improving diffusion strategies.)

Working backward from any detectable gains in the quality of undergraduate teaching, perhaps showing up in percentage acceptances at graduate schools, on Law School Acceptance Tests (LSAT) or other tests, on student ratings of courses and faculty, or on departmental examinations, we bring to bear the whole apparatus of MO analysis. We carefully examine causal lists, using enemies of the bureau as a vital source of alternative explanations (causes); we develop the most elaborate possible formulation of MO's for each of these candidates and then go after straight presence and MO completion checks; we set out explicitly the elements that discriminate between the members of this set of MO's, and set forth the evidence that we believe excludes the rejected candidates. We look with care at the span of the induction that is used to bring experience elsewhere or at other times to bear on this experiment, that is, at many of the inferences that generate our causal list. We check for co-causation and overdetermination. We use a goal-free or social process expert consultant to seek undesirable side effects in the academic jungle. We track down the sources of rumors and advice that are reported to have had significant effects, using journalists and historians as aides rather than educational psychologists and anthropologists; perhaps even a clinician could be helpful in digging into motives and memories to establish causal connections. We perform microexperiments to check the minipredictions that sometimes emerge from MO hypotheses. We may schedule a program of random (or random within the most probably influenced group) snapshots of the teaching process, looking for tracers or configural MO signatures. We should ultimately be able to develop a picture of the spreading influence of the bureau, in spite of its external and internal competitors, that is beyond the possibility of reasonable doubt. And perhaps we shall have narrowed the gap between respectable and anecdotal evaluation so that evaluation will seem a little less threatening to our humanistic friends because it is a little more familiar. Of course we have not handled all the problems for the evaluator. It must be realized that discovering:

A did cause X, where X is desirable,

does not always guarantee that A is desirable, although it justifies a prima facie inference to this conclusion. Of the many difficulties with this inference, some of which involve side effects, perhaps the strongest is the possibility that X could have been (or perhaps even would have been) achieved by a B that costs half as much as A. The second caution is also for the methodologist. The usual definitions of cause—explicit and implicit—are erroneous in a crucial respect, one where the MO method is not in error. Typical experimental designs for the investigation of psychotherapy success, for example, involve the assumption that parity of results between the experimental group and a no-treatment control group demonstrates that the psychotherapy is having no effect. This is an error because the controls may be engaged in autotherapeutic activities that simply have the same, significant-sized effect as psychotherapy. For an evaluator looking at interventions that might deserve federal support, this shows the psychotherapy to have zero cost-effectiveness, but it does not show zero effectiveness in the "scientific" sense. Either a second (pseudotherapy) control or MO analysis is required to distinguish this case from or identify it with a situation in which the effects of auto- and psychotherapy are both zero. Suppose those effects are not zero, only equal. Even though psychotherapy must be evaluated as a poor intervention strategy, there is a different framework in terms of which the evaluator would come to a different conclusion. Suppose the evaluator is looking at the options open to a neurotic patient. He or she will then rate (typically) the autotherapy and the psychotherapy as quite effective, and the latter as less cost-effective than the former. (Since the latter has a somewhat different MO from the former, including reduction of anxiety earlier and possibly serious withdrawal symptoms later, a time integration of the benefits may show some advantages for one or the other.) The question of "effectiveness" thus involves implicit reference to a base line, to a context of inquiry. This is more likely to occur to an evaluator, but it is true even for the scientist, who is less likely to accept the idea of relativity of effects.

One might argue that, for purposes of evaluation, one should always use the cruder notion of causation, which involves the axiom that the absence of the cause entails the absence of the effect, and hence the lemma that, if the absence of a putative cause leads to no difference in outcome compared to a group lacking the alleged cause, the "cause" has no effect, that is, it is not a cause. But the "insight function" of causal analysis is obviously blunted by such a usage, and, therefore, I reject the claim that it is mandatory.

Applying this point to the teaching bureau example, we have to realize that the discovery of many significant and desirable effects that really are due to its activities does not show that these effects would not have occurred if the bureau had never existed, perhaps with less cost. Causal power is not the same as utility or value; it is just a necessary condition for the latter.

Must MO methodology always be supported by control-group methodology? It depends on the question of interest. Is the question: Did the bureau really have a significant effect—or almost no effect? If so, then MO is appropriate. Or is the question one that is not usually distinguished from the first: Did the bureau bring about results that would not have occurred anyway? If this is the question, then you do supplementary work involving the usual controls or you simply have to make a bet based on intuition or trend analysis.

Don Campbell has suggested, in conversation, a neat way to set up the control group for this study if one gets in at the start of the intervention. One might ask the bureau to focus on the A-M faculty and use N-Z as control. But there are still real difficulties over the rejection of requests for help from the controls and leakage of the treatment. (It is always unfortunate to use an evaluation design where a real-world desirable effect has to be regarded as experimentally undesirable.) Once the system is off and running, there is no way to apply this design in retrospect. The MO approach can also be incorporated in the planning phase, as suggested, with decided advantages as far as getting a picture of the evaluee's total contribution is concerned. (For this reason, it is not just a subspecies of ex post facto design.) Although it will still work well as the study progresses, it will not tell you conclusively what would have happened if the treatment were not there. I think the evaluator or administrator must assume (an MO assumption) that external causes cannot be relied on to produce the desired effect either in regard to degree or to speed. The way to find out whether you have pulled the right lever is MO analysis.

The case we have been discussing involves a bureau that is given a strong interventionist charge (mission). Suppose it is given or adopts a low profile and simply responds to requests for help based on mere notification of its existence. Good results are somewhat more likely to be its very own and somewhat less likely to have been achieved by other factors in its absence. (On the other hand, it is less likely to achieve significant results in the absence of an aggressive campaign.) There is also a powerful credibility advantage if serious holistic evaluation is not done or if one's activities have all been based on requests for service and one can show MO evidence of services rendered in response to such requests. MO analysis of this sort can often show that the kind of assistance rendered simply was not available elsewhere, either when required or later. Nor would it have been available in the absence of the bureau because it would have required expertise and time only a central office could supply.

Campbell points out (in correspondence) some traps that bear on any serious classical evaluation of this kind of intervention. For example, in a controlled study one might find a highly significant correlation between treatment and *below*-average teaching performance, but this may be due to self-selection of treatment by the worst teachers. Or, if one looked at pretest-posttest gains, one might be seduced into thinking the treatment is highly successful with the low scorers on the pretest whereas the correlation with treatment is simply a consequence of a test-retest correlation. Of course there are analytical tools for protection against such misinterpretations (for example, using regressed pretest scores as the base line for gains in the second case). I am inclined to think that MO methodology is rather less liable to fall into such traps across the board. But we should examine more dangerous examples, where traps must be carefully avoided by the MO analyst.

Another context where this approach is called for is the accountability context: we need to know whether what was supposed to be done was done, and by the agency funded, but not necessarily whether it was done in the best possible way. Yet another context is the exploratory one in which we are investigating to see whether this independent variable can actually produce this effect, intending to do further studies if it can but to work elsewhere if it cannot. In areas where such studies are very costly in time, dollars, or skills, the MO approach provides a valuable early screening procedure.

Of great importance, it seems to me, is the fact that MO analysis does not violate the social-political taboos that sometimes render full experimental treatment impossible. It can be relatively unobtrusive.

One hope I have for this methodology is that it can help with Aptitude-Treatment-Interaction (ATI) problems. As is generally recognized, the confidence that significant ATI discoveries would emerge from the study of teaching styles and student characteristics such as cognitive style has so far been matched only by the barrenness of the results. An example that is of particular concern to me at the moment concerns role playing since I am evaluating a teacher-training package aimed at that goal. In defending it to me, the producer said that it did not really need any defense since he knew it worked "for some kids." (We could also take an example of large-scale social interventions like Head Start where the possibility exists that it worked very well at some sites.) I shall try to explain what I think the MO approach can do to test such hunches, which I believe are closely related to the *verstehen* intuitions of the historian and should not be dismissed as casually as is common.

The crudest approach to the evaluation of teaching role playing would perhaps consist in looking at the gain in the mean score on some appropriate instrument. It would be less crude to look at the distribution of individual gains. But this will not distinguish between two competing explanations of individual cases. Take a case like that Dave Berliner may have had in mind when speaking to me so favorably of role playing. The pre-post gain for a certain student has been very large, not only by comparison with other students in this class but also by "absolute" standards, that is, it is judged to represent a highly significant and difficult-to-achieve educational gain. We might throw in the fact that the child has greatly enjoyed the learning experience.

Now it is possible that this gain is simply a regression artifact, or it may be due to an educationally important matching of the treatment to the student characteristics. If we apply regression corrections to the scores, as Don Campbell was suggesting, we achieve the best statistically possible guide to the "genuineness" of the effect. But to do so will sometimes wrongly mask what is a real effect. Try to look at the situation with MO techniques in mind. Here is a student (or a school) with unusually low pretest scores. The specific meaning of "statistical artifact" here—or, I should say, the particular kind with which I am concerned at the moment and which seems to provide the most threatening competing explanation—is that the test items were an "unlucky" choice for this individual. There is a direct way of testing this, a retest, but it is not available after the treatment begins. It is also very costly, and there are hazards about its interpretation. One alternative explanation is that the individual's prior knowledge is really very limited. Now, are there differences in the way in which these two "causes" operate to produce their common effect, the low score? Both of them operate through the head of the subject, and that is where we should look for MO differentiae. Was it really the emptiness of the head that caused the low score? One obvious detector would be a tack-on question of a yes-no kind, which simply asks the student to rate his or her answers to this test as a fair indication of their present knowledge about the topic. In spite of the unreliability of such self-ratings, I am willing to bet there are populations of sufficient sophistication to make this a better indicator of "statistical artifact low-scorers" than anything else we have. (Remember, the students gain nothing by asserting that the test did not tap their knowledge. Their answer to this question can be deleted from their papers before correction to avoid contamination of the grader.)

Going a step further, can we design the test itself to discriminate between these causes? Of course, other things being equal, doubling its length is probably the best simple bet, but let us assume that is impossible. In any case, it is not optimal. Increasing the amount of choice of questions allowed the subject *is* desirable. Comparability of marking is thereby seriously weakened, but that is not usually fatal or we would not use item sampling. A good compromise, I have found, is a batch of compulsory items and then one or two open-ended, or almost open-ended, items like: "Discuss any other topic (or two topics) from this area or from the following list with which you are familiar." Even the use of multiple-choice questions in which any number of options, including zero, may be correct reveals a great deal more about the causes of a certain performance than the present ones.

It was a considerable step forward in testing to introduce item sampling, and it came about from rethinking the purpose of testing, realizing there is more than one, and seeing that the purposes would be best served by different kinds of tests. What I am suggesting is that MO analysis (and, indirectly, education) will be facilitated by certain types of tests currently in disfavor for the quite good but not conclusive reasons of lower scoring validity and more cumbersome processing.

Another MO check for ATI, perhaps more relevant for the role-playing example, would involve some process observations on each individual. These are expensive, but they are often desirable for other reasons. Look for *consonance,* for example, between (*a*) teacher's reports of quick learning or "surprising" lack of knowledge at an early stage and outstanding performance later, and (*b*) the pre's and the post's. We're looking for the MO signature of a highly effective teaching device, which will be absent if the effect is due to test unreliability.

In which of these cases should we use regressed pretest scores instead of raw scores? My instincts suggest looking at it both ways rather than standardizing on that correction. Experts may well know better, and I hope they will set the matter straight.

MO Corroboration

It was argued earlier that MO is not just a type of ex post facto design. Another reason for this view is the importance of combining MO checks with classical and semiclassical design. A nice example of where this would have helped was suggested in correspondence by Gene Glass after reading an earlier draft of this study. In *Pygmalion in the Classroom,* by R. Rosenthal and L. Jacobson (New York: Holt, Rinehart, and Winston, 1968), evidence is given from a classical design that teachers' beliefs about the potentialities of their students affect their estimates of the quality of work. Now the MO here is, of course, a causal chain involving the teachers' beliefs in the existence of differences between students that in some cases did not exist (since the teachers had been deliberately misinformed that some students were latently outstanding). An MO check would involve, inter alia, checking for the presence of each key cognitive component at the time when it was supposed to be functioning. Interestingly enough, there was a check on one such item, namely, the teachers' memory of which students were, according to their information, latent high performers. It turned out that the teachers could not remember to any significant degree which students were in this category. This automatically makes the study's overall conclusion very implausible; psychological causes cannot operate across time without intervening links. One must explore the possibility of unconscious retention, but it is here extremely implausible. Hence, one is not surprised to find that the statistics, which led to the apparent significance of the results, were faulty. (See Rosenthal and Jacobson, *Pygmalion in the Classroom*; the best summary of critiques appears in *Second Handbook of Research on Teaching,* ed. R. M. W. Travers [Chicago: Rand McNally, 1973], especially T. X. Barber, "Pitfalls in Research," on pages 393-398.)

In designing any classical study, it generally seems worthwhile to devote a little time to the questions: What is the means whereby the putative cause is supposed by bringing about the effect? What are the links in the causal chain between them? Can we look for these links or arrange that they will be easy to look for? Can we use their occurrence to distinguish between the alternative causal hypotheses? How?

Further Directions

There are a number of directions in which I would like to expand this discussion if I had space, time, and talent. Let me conclude by mentioning them briefly in the hope that they will start others thinking or talking or even both.

1. I would like to show how MO technique applies in the field of economic analysis, especially economic case studies, for in that field control groups are essentially impossible and MO technique is correspondingly important.
2. I would like to show how "microexperiments" to test MO hypotheses are often possible where full experimental studies (of the whole treatment) are impossible.
3. I would like to discuss the various species of overdeterminism, some of which can only be "resolved" by MO analysis.
4. I would like to point out the fallacy inherent in the statistician's incorrect use of the term "explained" when referring to distributing the variance amongst various factors; MO analysis clarifies it.
5. And, finally, I would like to explore the quantitative dimension of MO analysis, the ways in which we could convert it into pattern-recognition computer programs and make it amenable to covariance analysis.

Common sense leads us to develop the science of using control groups and randomization, and, when this fails, it leads us to quasi-experiments. Even when these are impossible, we have not exhausted the resources of common sense. MO methodology can be used both as an alternative to, and a strengthening of, the other approaches. It gives us greater robustness and better defenses against the assaults of attrition and artifacts.

Note: The inference drawn on pages 69-71 is not deductively valid, only mnemonically useful. It needs one—or a set—of other premises to yield a strong probability inference; *which* ones we can use depends on the particular kind of phenomenon. For example, it is typically the case that the set A_i includes all the "big" candidates; formally this means that, if the unknown causes are B_i, then the relative frequency of any A_i in the class of X's is much greater than that of any B_i (roughly because we would have identified any big B_i). Hence, if the only big candidate that is present is A, its only competitors are low-frequency events, and A is thus *the most likely* of all the possible causes, i.e., max P(A_i caused X) = P(A caused X). For it to be *most likely the cause,* i.e., for P(A caused X) $>$.5, we need a further condition, for example, that the frequency of A's (in the class of X-occurrences) is greater than the sum of the frequencies of all B_i. This is the relevant sense of "almost all." This condition is more easily met the more evenly the A_i divide the spoils, the smaller the number of big candidates is, and the smaller the proportion of unexplained cases. For example, if three, four, five, . . . , up to nine known causes account for 91 percent of X's, then (as long as each of them accounts for at least 10 percent of X's) the inference pattern works. But if only four causes are known to account for only 80 percent of X's, we cannot pull off the inference to any of them as a better-than-even chance. Still—often importantly—we can get the "modal" conclusion (that A is the best bet to be the cause) as long as the weaker assumption holds (that no B_i is as often successful as any A_i). (Many thanks to Frederick Mosteller for identifying and forcing me to bridge the logical gap here.)

COST ANALYSIS IN EVALUATION AND THE DOCTRINE OF COST-FREE EVALUATION

Costs are shadowy figures hovering in the background of evaluation, specters that come to haunt those who try to ignore them, Janus-faced figures more elusive than the most ghostly of the mental entities to which the hard-nosed empiricist objects in the scientific context, and yet the very substance of the hard-nosed empiricist's position in the management area.

I wish to set out some of the conceptual and practical difficulties that currently concern me in the hope that others more expert than I—economists or accountants, perhaps—can provide answers. For the evaluator, such answers are badly needed, indeed; where I have ventured one, it is with a strong sense of my own limitations in this kind of analysis and mainly in the hope that providing a target to shoot at may improve the aim of the expert. It is less easy to simply take the advice of experts in this area than it is in many areas for two reasons. First, the way that experts use the notion of cost is often highly technical and, although the advantages of such an approach are probably considerable in the expert's special field, for example, cost accounting, it is not at all obvious that the educator, author, federal agency, legislator, or evaluator will gain from adopting the technical use. Second, the interjudge reliability of the experts in the crunch is not impressive, as Abraham Briloff (*Unaccountable Accounting* [New York: Harper and Row, 1972]) so ably documents. There may be some value in starting afresh to work out the conceptual points for oneself. I begin with a few elementary facts and examples in order to lay the ground for others. Even these examples suggest that the usual level of consideration is pretty superficial.

Direct and Indirect Costs

You tell your secretary to order a book through the mail, sending along a check with the order. The book "costs" $7.50. To whom? Not to you. Its cost to you is $7.50 plus the cost of your time spent dictating the order plus the direct cost of your secretary's time transcribing, typing, filing, checkwriting, folding, stamping, and mailing (better than $2.50 on the usual estimates), plus postage, plus proportional amortization of office equipment and capital and maintenance costs, plus a slice of rent, heating, cleaning, mortgage interest, insur-