*Annu. Rev. Psychol. 1994. 45:545–80*

# SOCIAL EXPERIMENTS: SOME DEVELOPMENTS OVER THE PAST FIFTEEN YEARS

*Thomas D. Cook*

Departments of Sociology and Psychology, Northwestern University, Evanston, Illinois 60208

*William R. Shadish*

Department of Psychology, Memphis State University, Memphis, Tennessee 38152

KEY WORDS: causation, randomization, quasi-experiments, causal generalization, selection modeling

CONTENTS

# INTRODUCTION

The *Annual Review of Psychology* has never published a chapter on causal inference or experimentation. This is surprising, given psychology's traditional concern with establishing causal relationships, preferably through experiments. Our own expertise is in social experiments, as practiced not only in psychology but also in education, economics, health, sociology, political science, law and social welfare. Philosophers and statisticians are also involved with such experiments, but more to analyze them than to do them. This review concentrates on important developments from this multidisciplinary spectrum during the 1980s and early 1990s.

Experiments can be characterized both structurally and functionally (Cook 1991a). The more prototypical structural attributes of experiments include a sudden intervention; knowledge of when the intervention occurred; one or more post-intervention outcome measures; and some form of a causal counterfactual—that is, a baseline against which to compare treated subjects. Functionally, experiments test propositions about whether the intervention or interventions under test are causally responsible for a particular outcome change in the restricted sense that the change would not have occurred without the intervention.

Social experiments take place outside of laboratories and therefore tend to have less physical isolation of materials, less procedural standardization, and longer-lasting treatments when compared to experiments in laboratory settings. Social experiments are usually designed to test an intervention or treatment that is usually better characterized as a global package of many components than as a presumptively unidimensional theory-derived causal construct. This is because the usual aim of field researchers is to learn how to modify behavior that has proven to be recalcitrant in the past (e.g. poor school performance, drug abuse, unemployment, or unhealthful lifestyles) as opposed to testing a theoretical proposition about unidimensional causes and effects. There are two types of social experiments, each of which has the sudden intervention, knowledge of intervention onset, posttest, and causal counterfactual component that characterize all experiments. Randomized experiments have units that are assigned to treatments or conditions using procedures that mimic a lottery, whereas quasi-experiments involve treatments that are assigned nonrandomly, mostly because the units under study—usually individuals, work groups, schools or neighborhoods—self-select themselves into treatments or are so assigned by administrators based on analysis of who merits or needs the opportunity being tested.

We cover four experimentation topics that are currently of interest: 1. modifications to the dominant theory of social experimentation; 2. shifts in thinking about the desirability and feasibility of conducting randomized field experiments; 3. modifications to thinking about quasi-experiments; and 4. recent concerns about justifying generalized causal inferences.

## CHANGES IN THE DOMINANT THEORY OF SOCIAL EXPERIMENTATION

### *Theories of Causation*

Experimentation is predicated on the manipulability or activity theory of causation (Collingwood 1940, Mackie 1974, Gasking 1955, Whitbeck 1977), which tries to identify agents that are under human control and can be manipulated to bring about desired changes. This utilitarian conception of causation corresponds closely with common-sense and evolutionary biological understandings of causation, and it assumes that cause-effect connections of this kind are dependable enough to be useful. Experiments have this same purpose. They probe whether a force that is suddenly introduced into an ongoing system influences particular outcomes in a way that would not have occurred without the intervention. The aim of experiments is to describe these causal consequences, not to explain how or why they occurred.

Experiments can be made more explanatory, though. This is achieved primarily by selecting independent and dependent variables that explicitly explore a particular theoretical issue or by collecting measures of mediating processes that occurred after a cause was manipulated and without which the effect would presumably not have come about. But few social experiments of the last decades were designed to have primarily an explanatory yield, and nothing about the logic of experimentation requires a substantive theory or well-articulated links between the intervention and outcome, though social experiments are superior if either of these conditions is met (Lipsey 1993).

Understanding experiments as tests of the causal consequences of manipulated events highlights their similarity with the manipulability theory of causation. But it also makes them seem less relevant to the essentialist theories of causation (Cook & Campbell 1979) to which most scholars subscribe. These theories seek either a full explanation of why a descriptive causal connection comes about or the total prediction of an effect. Thus, they prioritize on causal explanation rather than causal description, on isolating why a causal connection comes about rather than inferring that a cause and effect are related. Such explanatory theories are likely to be reductionistic or involve specifying multiple variables that co-condition when a cause and effect are related. Each of these is a far cry from the "if X, then Y" of the manipulability theory of causation. Mackie (1974)—perhaps the foremost theorist of causation—sees the manipulability theory as too simplistic because it assumes a real world characterized by many main effects that experiments try to identify. For

Mackie and the essentialists, the real world is more complex than this. However, Mackie doubts whether essentialist theories are epistemologically adequate, however ontologically appropriate they may be. He postulates that all causal knowledge is inevitably "gappy," dependent on many factors so complexly ordered that full determination of the causal system influencing an outcome is impossible in theory and practice. Thus, he contends that essentialists seek a complete deterministic knowledge they are fated never to attain.

Conceiving of all causal relationships as embedded within a complex context that co-determines when an effect is manifest implies the need for a methodology sensitive to that complexity and embeddedness. Yet individual experiments are designed to test the effects of one or a few manipulable independent variables over a restricted range of treatment variants, outcome realizations, populations of persons, types of settings, and historical time periods. To be sure, in their individual studies, researchers can measure attributes of treatments, outcomes, respondents, and settings and then use these measures to probe whether a particular outcome depends on the treatment's statistical interaction with such attributes. This strategy has been widely advocated, especially in education (Cronbach & Snow 1976). But the variability available for analysis in individual studies is inherently limited, and tests of higher-order interactions can easily lead to data analyses with low statistical power (Aiken & West 1991). As a result, interest has shifted away from exploring interactions within individual experiments and toward reviewing the results of many related experiments that are heterogeneous in the times, populations, settings, and treatment and outcome variants examined. Such reviews promise to identify more of the causal contingencies implied by Mackie's fallibilist, probabilistic theory of causation; and they speak to the more general concern with causal generalization that emerged in the 1980s and 1990s (e.g. Cronbach 1982, Cook 1993, Dunn 1982).

With its emphasis on identifying main effects of manipulated treatments, the activity theory of causation is too simple to reflect current philosophical thinking about the highly conditional ways in which causes and effects are structurally related. The experimentalist's pragmatic assumption has to be either (*a*) that some main effects emerge often enough to be dependable, even if they are embedded within a larger explanatory system that is not fully known; or (*b*) that the results of individual experiments will eventually contribute to a literature review identifying some of the causal contingencies within this system.

### *Theories of Categorization*

The descriptive causal questions that experimenters seek to answer usually specify particular cause and effect constructs or categories to which generalization is sought from the manipulations or measures actually used in a study.

Philosophical work on categorization is currently dormant, while in psychology there is considerable productive chaos. Classic theories of categorization postulate that instances belong in a class if they have all the fixed attributes of that class. This approach has been widely discredited (Rosch 1978, Lakoff 1987, Medin 1989), and there is now recognition that all categories have fuzzy rather than clearly demarcated boundaries (Zimmerman et al 1984), that some elements are more central or prototypical than others (Smith & Medin 1981), and that knowledge of common theoretical origins is sometimes more useful for classifying instances than the degree of initially observed similarity (Medin 1989). In biology, plants or animals used to be assigned to the same class if they shared certain attributes that facilitate procreation. But classification now depends more on similarity in DNA structure, so that plants or animals with seemingly incompatible reproductive features but with similar DNA, are placed in the same category. This historical change suggests that seeking attributes that definitively determine category membership is a chimera. The attributes used for classification evolve with advances elsewhere in a scholarly field, as the move from a Linnean to a more microbiological classification system illustrates in biology.

Categorization is not just a logical process; it is also a complex psychological process (or set of processes) that we do not yet fully understand. Ideally, social experiments require researchers to explicate a cause or an effect construct, to identify its more prototypical attributes, and then to construct an intervention or outcome that seems to be an exemplar of that construct. Explicit here is a pattern-matching methodology (Campbell 1966) based on proximal similarity (Campbell 1986). That is, the operation looks like what it is meant to represent. But the meanings attached to similar events are often context-dependent. Thus, if a study's independent variable is a family income guarantee of $20,000 per year, this would probably mean something quite different if the guarantee were for three years versus twenty. In the first case individuals would have to think much more about leaving a disliked job because they do not have the same financial cushion as someone promised the guarantee for 20 years. Also, classifying two things together because they share many similar attributes does not imply they are theoretically similar. Dolphins and sharks have much in common. They live exclusively in water, swim, have fins, are streamlined, etc. But dolphins breathe air out of water and give direct birth to their young—both features considered prototypical of mammals. Hence, dolphins are currently classified as mammals rather than fish. But why should breathing out of water and giving birth to one's young be considered prototypical attributes of mammals? Why should it not be the fit between form and function that unites dolphins and fish? Another example of this complexity comes from social psychology. Receiving very little money to do something one doesn't want to do seems different on the surface from being asked to choose between two objects of similar value. Yet each is supposed to elicit the same theoretical process of cognitive dissonance. Should the attributes determining category membership depend on physical observables linked to particular classes by a theory of proximal exemplification or prototypicality? Or should the attributes instead reflect more latent and hence distal theoretical processes, such as those presumed to distinguish mammals from fish or those that make financial underpayment and selecting among objects of similar value instances of the same class called cognitive dissonance?

Experimenters have to generalize about causes, effects, times, settings, and people. If the best theories from philosophy and psychology are not definitely helpful, researchers can turn instead to statistics where formal sampling theory provides a defensible rationale for generalization. However, in experimental practice it is almost impossible to sample treatment variants, outcome measures, or historical times in the random fashion that sampling theory requires; and it is extremely difficult to sample persons and settings in this way. So, in individual experiments, generalization depends on the purposive rather than random selection of instances and samples to represent particular categories of treatment, outcome, people, or settings. Unfortunately, in current sampling theory, purposive selection cannot justify generalization to a category—only random selection can.

## *Relabeling Internal Validity*

Among social scientists, Campbell (1957) was the first to systematically work through the special issues that arise when probing causal hypotheses in complex field settings. As he recounts it (Campbell 1986), his analysis originated from concerns with the low quality of causal inferences promoted by the field research of the time and with the questionable generalizability of the cause-testing laboratory research of the time. So he invented a language for promoting more confident causal inferences in contexts that better resembled those to which generalization is typically sought. He coined the term *internal validity* to refer to inferences about whether the relationship between two operationalized variables was causal in the particular contexts where the relationship had been tested to date; and he invented *external validity* to express the extent to which a causal relationship can be generalized across different types of settings, persons, times, and ways of operationalizing a cause and an effect. These two terms have now entered into the lexicon of most social scientists.

But many scholars did not understand internal validity in the way Campbell intended. It was widely taken to refer, not just to the quality of evidence about whether an obtained relationship between two variables was causal, but also to whether the relationship was from a particular theoretically specified causal agent to a particular theoretically specified effect (Kruglanski & Kroy 1975, Gadenne 1976). Thus, labeling the cause and effect were smuggled in as components of internal validity. Cronbach (1982) went even further, adding that internal validity concerns whether a causal relationship occurs in particular identifiable classes of settings and with particular identifiable populations of people—each an element of Campbell's external validity. To discuss validity matters exactly, Cronbach invented a notational system. He used *utos,* to refer to the instances or samples of units *u,* treatments *t,* observations *o,* and settings *s* achieved in a research project probing a causal hypothesis. He used *UTOS,*[1] to refer to the categories, populations, or universes that these instances or samples represent and about which research conclusions are eventually drawn. Translating Campbell's internal validity into these terms we see that it refers only to the link between *t* and *o* at the operational level. What *t* stands for (i.e. how the cause or *T* should be labeled) is irrelevant, just as it is irrelevant what *o* stands for (i.e. how the effect or *O* should be labeled). For Campbell, these are matters of external validity, as are concerns about the types of units and settings in which the *t-o* relationship can be found (Campbell 1957, Campbell & Stanley 1963).

After critics eventually understood how restricted Campbell's description of internal validity was, they decried it as positivistic (Gadenne 1976), irrelevant to substantive theory (Kruglanski & Kroy 1975), and insensitive to the contingency-laden character of all causal conclusions (Cronbach 1982). Cronbach persisted in describing his internal validity in terms of generalization from all parts of *utos* to all parts of *UTOS,* thus subsuming under his version of internal validity all the elements Campbell had included as elements of his external validity! To reduce this ambiguity, Campbell (1986) sought to invent more accurate labels for internal validity. He invented *local molar causal validity* to replace internal validity, hoping to highlight with *local* that internal validity is confined to the particular contexts sampled in a given study. By *molar* he hoped to emphasize that his internal validity treated interventions and outcomes as complex operational packages composed of multiple microelements or components. (The contrast here is with more reductionist approaches to causation that place most weight on identifying the microelements responsible for any relationship between a more molar treatment and a more molar outcome.) Finally, with the *causal* part of his new label, Campbell sought to highlight the centrality of assertions about whether the *o* in Cronbach's *utos* formulation would change without variation in *t,* thereby downplaying the aspiration to learn about other facets of this link, especially its generalizability. Campbell's new name for internal validity has not been widely adopted, despite its greater precision. Internal validity is still widely

[1] We ignore here Cronbach's discussion of the difficulty of drawing inferences about populations or settings.

used and misused, though clarification is now at hand for those willing to seek it.

## *Relabeling External Validity*

Campbell's concept of external validity also came to be modified. His early work described external validity in terms of identifying the contexts in which a given causal relationship holds, including ways that the cause or effect were operationalized. Cook & Campbell (1979) distinguished between generalizing to theoretical cause and effect constructs and generalizing to populations of persons and settings, using construct validity to refer to the former and external validity to refer to the latter. More importantly, Cronbach (1982) differentiated between two types of generalization that were obscured in the work of Campbell and his colleagues. The first involves generalizing from *utos* to *UTOS*—using attributes of the sampling particulars to infer the constructs and populations they represent. This is what Cronbach calls internal validity. The second involves generalizing from *UTOS* to **UTOS* (star-*UTOS*), with **UTOS* representing populations and constructs that have manifestly different attributes from those observed in the sampling details of a study. At issue with **UTOS* is extrapolation beyond the sampled particulars to draw conclusions about novel constructs and populations. Cronbach calls this external validity.

To clarify his own different understanding of external validity, Campbell relabeled it as *proximal similarity* (Campbell 1986). He chose this label to emphasize a theory of generalization based on using observable attributes to infer whether a particular sample of units, treatments, observations, or settings belonged in a particular class or category as that class or category is usually understood within a local research community or in ordinary English-language usage. Similarity between the specific instances sampled and the more general classes they are supposed to represent is the order of the day, and Campbell is adamant in classifying instances on the basis of observable characteristics linked to a theory specifying which characteristics are prototypical. *Proximal* is part of his new label for external validity to emphasize that other characteristics are more distal for him, particularly those where the similarity is not immediately evident but is instead derived from substantive theory—e.g. classifying dolphins as mammals rather than fish. Like local molar causal validity, proximal similarity has failed to catch on among social scientists. Campbell's old external validity lives on, seemingly as vigorous as ever. However, the need he felt in 1986 to create new labels to clarify the constructs first introduced in 1957 suggests growing dissatisfaction about the clarity or content of his validity types. However worthy the underlying ideas, the terms themselves cannot now be considered as unproblematic and as sacrosanct as they once were.

## *Debates about Priority among Validity Types*

One debate about priorities among validity types concerns the warrant for Campbell's preference for internal over external validity. Campbell & Stanley (1963) argued that it is only useful to explore how generalized a causal connection is if very little uncertainty remains about that connection. How would theory or practice be advanced, they reasoned, if we identified limits to the generalization of a relationship that might later turn out not to be causal? Cronbach (1982) asserted that this priority reflects the conservative standards scientists have traditionally adopted for inferring causation, and he questioned whether these standards are as relevant for other groups as they are for individuals with a stake in the results of a social experiment. His own experience has convinced him that many members of the policy-shaping community are more prepared than scholars to tolerate uncertainty about whether a relationship is causal, and they are willing to put greater emphasis than scholars do on identifying the range of application of relationships that are manifestly still provisional. Cronbach believes that Campbell's preference for internal over external validity is parochial. Chelimsky (1987) has offered an empirical refutation of Cronbach's characterization of knowledge preferences at the federal level. But research on preferences at the state or local levels is not yet available, and Cronbach's characterization may be more appropriate there. Time will tell.

Dunn (1982) has criticized the validity typology of Campbell and his colleagues because neither internal nor external validity refers to the importance of the research questions being addressed in an experiment. He claims that substantive importance can be explained just as systematically as the more technical topics Campbell has addressed, and he has proposed some specific threats that limit the importance of research questions. Dunn is correct in assuming that there is no limit to the number of possible validity types, as Cook & Campbell partially illustrated when they divided Campbell's earlier internal validity into statistical conclusion validity (are the cause and effect related?) and internal validity (is the relationship causal?), and when they divided his external validity into construct validity (generalizing from the research operations to cause and effect constructs) and external validity (generalizing to populations of persons and settings). Cook & Campbell also proposed new threats to validity under each of these headings, and their validity typology is widely used (e.g. Wortman 1993). Thus, Campbell and his colleagues agree with Dunn, only criticizing him on some particulars about threats to the importance of research questions (see especially Campbell 1982).

Though his higher-order validity types came under attack, Campbell's specific list of threats to internal and external validity did not. Internal validity threats like selection, history, and maturation continue to be gainfully and regularly used to justify causal inferences, under whichever validity label they might be fit. This is probably because researchers, in their daily work, have to rule out specific threats to particular inferences. Debates about validity types are more abstract and remote. Campbell's identification of specific validity threats constitutes one of his most enduring contributions to social science methodology.

## *Raising the Salience of Treatment Implementation Issues*

Cook & Campbell (1979) distinguished between construct and external validity because (*a*) the means for promoting construct validity are rarely, if ever, based on the random sampling procedures so often advocated for generalizing to populations of persons and settings; and (*b*) experience has indicated how problematic treatment implementation can be in nonlaboratory contexts where researchers are more often guests than hosts, where researchers are not always the providers of services, where program theory is often poorly explicated, and where unexpected difficulties often arise to compromise how well an intervention is mounted (Boruch & Gomez 1977; Sechrest et al 1979; Bickman 1987, 1990; Chen & Rossi 1983; Gottfredson 1978). Hence, Cook & Campbell (1979) outlined a new list of threats describing some of the ways the complex intervention packages implemented in practice differ from the often more theoretical intervention procedures outlined in a research protocol (Lipsey 1993). Their concern was to identify which parts of the research protocol were actually implemented; which were not implemented; which were implemented but without the frequency or intensity detailed in the guiding theory; and which unplanned irrelevancies impinged on the research and might have influenced research outcomes.

Experiments test treatment contrasts rather than single treatments (Holland 1986). Cook & Campbell (1979) drew attention to four novel threats that affect the treatment contrast without necessarily influencing the major treatment purportedly under test. All these threats touch on diffusion of one treatment to other treatments. *Resentful demoralization* occurs when members of control groups or groups receiving less desirable treatments learn that other groups are receiving more. If they become resentful because of this focused comparison their performance might decrease and lead to a treatment-control difference. But this is because the controls are getting worse rather than the experimentals getting better. They also postulated that *compensatory rivalry* can arise if the focused inequity that so many experimental contrasts require leads members of the groups receiving less to respond by trying harder to show that they do not deserve their implicitly lower status. Such a motivation obscures true effects occurring in treatment groups because performance in the control groups should improve. *Compensatory equalization* occurs when administrators are not willing to tolerate focused inequities and they use whatever discretionary resources they control to equalize what each group receives. This also threatens to obscure true effects of a treatment. Finally, *treatment diffusion* occurs when treatment providers or recipients learn what other treatment groups are doing and, impressed by the new practices, copy them. Once again, this obscures planned treatment contrasts. These four threats have all been observed in social experiments, and each threatens to bias treatment effect estimates when compared to situations where the experimental units cannot communicate about the different treatments. Statisticians now subsume them under the general rubric of SUTVA—the stable-unit-treatment-value assumption (Holland & Rubin 1988)—in order to highlight how much the interpretation of experimental results depends on the unique components of one treatment group not diffusing to other groups and hence contributing to the misidentification of the causal agent operative within a treatment contrast. Is it the planned treatment that influences the outcome, or is it serving in a control group?

This is only one of many arguments in favor of documenting the quality of treatment implementation in all treatment conditions, including no-treatment controls. "Black box" experiments without such documentation are potentially counterproductive (Lipsey 1993). If they result in no-difference findings, it is not clear whether the intervention theory was wrong, the particular treatment implementation was weak, or the statistical power to detect a given effect was inadequate. In the absence of documentation of treatment integrity, even positive results do not make it clear what role the intended processes played in bringing about outcome changes. Realizing this, experimenters have devoted more resources over the last decade to describing implementation quality and relating this to variability in outcomes.

## *The Epistemology of Ruling Out Threats*

All lists of validity threats are the product of historical analyses aimed at identifying potential sources of bias. They are designed to help researchers construct the argument that none of these threats could have accounted for a particular finding they wish to claim as causal. The epistemological premise here is heavily influenced by Popper's falsificationism (1959). But Kuhn's (1962) critique of Popper on the grounds that theories are incommensurable and observations are theory-laden makes it difficult to believe that ruling out such threats is easy. So, too, does Mark's (1986) observation that ruling out validity threats often depends on accepting the null hypothesis—a logically tenuous exercise. Campbell has always emphasized that only plausible threats need ruling out and that plausibility is a slippery concept, adding to the impression that falsification is logically underjustified.

However, in the real world of research there have to be some mechanisms for certifying which threats are plausible in local contexts and for establishing which procedures will probably rule out a particular threat. The scientific community is not likely to prefer methods for ruling out threats that depend only on professional consensus. Logically justifiable procedures are the preference. This is why causal inferences from quasi-experiments are less convincing than from randomized experiments. They depend on a larger set of accompanying arguments, including many that are weaker because they postulate that particular threats are implausible in the context under analysis or have been ruled out on the basis of statistical adjustments of the data that themselves are assumption-laden. Is it not better to control threats via random assignment, with its clear rationale in logic? Realization of the epistemological difficulties inherent in ruling out alternative interpretations has strengthened calls both to sample persons, settings, treatments, and outcomes at random and also to assign units to treatments at random (Kish 1987). There is no debate about the desirability of this in principle; however, there are legitimate debates about the feasibility of random assignment and random selection when testing causal hypotheses, and especially about the degree of confidence merited by different methods that do not use some form of randomization for inferring cause and estimating its generalizability.

## RANDOMIZED EXPERIMENTS IN FIELD SETTINGS

The strength of the randomized experiment seems to be that the expected mean difference between randomly created groups is zero. This is merely a special case of the more general principle that unbiased causal inference results whenever all the differences between groups receiving treatments are fully known (Rubin 1977, Trochim 1984). Statisticians like to remind us that unbiased causal inference results if individuals are assigned to conditions by the first letter of their surname, by whether they are left or right handed, or by the order in which they applied to be in a particular social program. The key is that the assignment process is fully known and perfectly implemented; random assignment is only one instance of this more general principle.

Though random assignment involves considerable ethical, political, and legal problems (detailed in Coyle et al 1991), assertions that such assignment can rarely be used outside the laboratory have been disproved by the long lists of experiments Boruch et al (1978) have compiled. Cronbach (1982) has reclassified some of the studies on these lists as quasi-experiments or failed randomized experiments; nonetheless, the lists leave the impression that random assignment is already common in social science and could be even more widely used if researchers were better informed about how to circumvent objections to it. Researchers wanting to use other means of assignment now find it more difficult to justify their plans than even a decade ago. There are still cases where a randomly created control group is impossible (e.g. when studying the effects of a natural disaster), but the last decade has witnessed a powerful shift in scientific opinion toward randomized field experiments and away from quasi-experiments or non-experiments. Indeed, Campbell & Boruch (1975) regret the influence Campbell's earlier work had in justifying quasi-experiments where randomized experiments might have been possible. In labor economics, many critics now advocate randomized experiments over quasi-experiments (Ashenfelter & Card 1985, Betsey et al 1985, LaLonde 1986). In community-based health promotion research, randomized experiments have now replaced the quasi-experiments that were formerly used to reduce the high cost of mounting interventions with the larger sample size of whole towns or cities that random assignment requires. Indeed, in biostatistics much emotional strength is attached to random assignment and alternatives are regularly denigrated (e.g. Freedman 1987).

Much of the recent discussion about randomized experiments deals with implementing them better and more often. Little of this discussion has appeared in scholarly outlets. So the following exposition of the factors making randomized experiments easier to implement depends heavily on informal knowledge among practicing social experimenters as well as on such published sources as Boruch & Wothke (1985).

### *Mounting a Randomized Experiment*

The best situation for random assignment is when the demand for the treatment under evaluation exceeds the supply. The treatment then has to be rationed and can be allocated at random. If demand does not exceed supply naturally, experimenters can often induce the extra demand by publicizing the treatment. In a pilot study of multiracial residential housing in a university, CM Steele (unpublished research proposal) found in the pilot year that too few whites volunteered for the program to make random assignment possible. So, the next year he publicized the residential units through the campus housing office. Such publicity creates a tradeoff. It makes random assignment easier; but because of it, those who eventually enter the control or placebo conditions have learned of the multiracial housing alternative for which they applied but were not selected. This knowledge might lead to treatment diffusion, compensatory rivalry, compensatory equalization, or resentful demoralization. Although such threats can operate in any type of study, the question for field experimenters is whether they operate more strongly because of the publicity required to stimulate extra demand (Manski & Garfinkel 1992) or because random assignment provides a less acceptable rationale for experimental inequities when compared to allocation by need, by merit, or on a "first come, first served" basis (Cook & Campbell 1979).

When using random assignment in field settings, a key concern is when respondents are assigned to treatments. Riecken & Boruch (1974) identified three options: before respondents learn of the measurement burdens and treatment alternatives; after they learn of the measurement demands, but before they learn of the content of the treatments being contrasted; and after they know of the measurement demands and treatments and agree to be in whatever treatment condition the coin toss (or equivalent thereof) determines for them. Opinion has long favored the last alternative because it reduces the likelihood that refusals to serve in the experiment will be related to the treatments, thus creating the nonequivalence that random assignment is designed to avoid. However, the cost to such delayed randomization is that some persons may drop out of the study when they realize they are not guaranteed the treatment they most desire. Thus, the potential gain in internal validity (in Campbell's original sense) entails a potential loss in external validity.

Some scholars now question the wisdom of placing such a high premium on internal validity that treatment assignment is delayed. For policy contexts, Heckman & Hotz (1988) question the usefulness of strong causal conclusions if they only generalize to those who are indifferent about the treatments they receive, because personal belief in a treatment might be an important contributor to its effectiveness. This critique highlights the advantages that follow from linking a randomized study of persons willing to undergo any treatment with a quasi-experimental study of persons who self-select themselves into treatments, controlling for self-selection in the quasi-experiment as well as the statistical state of the art allows (Boruch 1975). Otherwise, advocates of delayed random assignment have to argue that it is better to have limited information about the generalization of a confident causal connection than it is to have considerable information about the generalizability of a less certain connection.

Using random assignment in complex field settings is all the more difficult because researchers often do not do the physical assignment process themselves. They may design the process and write the protocols determining assignment, whether as manuals or computer programs implementing a complex stratification design. But the final step of this assignment process is often carried out by a social worker, nurse, physician, or school district official. It is up to these profesionals to explain to potential volunteers the rationale for random assignment, to detail the alternatives available, and then to make the actual treatment assignment. Sometimes, implementers at the point of service delivery misunderstand what they are supposed to do—which is mostly a matter of improved training. At other times, however, professional judgment predominates over the planned selection criteria and assignment occurs by presumptions about a client's need or merit instead of by lottery.

Professionals are trained to judge who needs what, when. It is not always easy for them to let a manual or computer protocol decide. Indeed, we have heard anecdotes about workers in early childhood development centers who came in surreptitiously at night to modify a computer so that it made the assignments they believed were professionally correct rather than the random assignments that were supposed to be made! To prevent professional judgment from overriding scientific assignment requires giving better explanations of the necessity for random assignment, having local research staff carry out the assignment rather than service professionals, instituting earlier and more rigorous monitoring of the selection process, and providing professionals with the chance to discuss special cases they think should be excluded from the assignment process.

Random assignment can be complicated, especially when social programs are being tested. In such cases, the organization providing benefits could lose income if all the places in the program are not filled. This means a waiting list must be kept; however, it is unrealistic to expect that all those on the waiting list will suspend self-help activities in the hope they will eventually enter the program. Some will join other programs with similar objectives or will have their needs met more informally. Thus, even a randomly formed waiting list changes over time, often in different ways from the changes occurring in a treatment group as it experiences attrition. Though no bias occurs if persons from the waiting list are randomly assigned to treatment and control status when program vacancies occur, it can be difficult to describe the population that results when people from the ever-changing waiting list are added to the original treatment group. Bias only occurs with more complex stratification designs where program needs call for making the next treatment available to a certain category of respondent (e.g. black males under 25). If there are fewer of these in the pool than there are treatment conditions in the research design, then program needs dictate that this person enters the program even if no similar person can be assigned to the control status. If researchers belatedly recruit someone with these characteristics into a comparison group, this vitiates random assignment and introduces a possible selection bias. But if they exclude these persons from the data analysis because there are no comparable controls, this reduces external validity.

Outside the laboratory, random assignment often involves aggregates larger than individuals (e.g. schools, neighborhoods, cities, or cohorts of trainees). This used to present problems of data analysis since analysis at the individual level fails to account for the effects of grouping, while analysis at the higher level entails smaller sample sizes and reduced statistical power. The advent of hierarchical linear modeling (Bryk & Raudenbush 1992) should help reduce this problem, particularly when more user-friendly computer programs are available. But such modeling cannot deal with the nonequivalence that results when, despite random assignment, the number of aggregated units is low. After all, it can be costly to assign whole cities to treatments and obtain stable estimates for each city at each measurement wave. To counter this, researchers now almost routinely match larger units on pretest means or powerful correlates of the outcome before assigning them to the various treatments. Since the highest single research cost is likely to be mounting the intervention, researchers sometimes select more control than intervention sites (Kish 1987). This reduces the statistical power problem associated with small sample sizes, but it does not guarantee initial group equivalence if the treatment is still assigned to a small sample of randomly selected and aggregated units. So the preferred solution is to move to a study with more schools or cities. The National Heart, Lung and Blood Institute now seems to have adopted this point of view, preferring multimillion dollar, multisite studies of smoking cessation and women's health over smaller studies. Yet these studies will inevitably be homogeneous in some components that reduce generalization, and critics will inevitably suggest some study factors that need to be varied in subsequent research. Moreover, the expense of such multisite experiments entails opportunity costs—meritorious studies that deserve to be funded cannot be.

## *Maintaining a Randomized Experiment*

Unlike laboratory experimenters, field experimenters are usually guests in somebody else's organization. The persons they study are usually free to come and go and pay attention or not as they please, lowering the degree of treatment standardization. Some individuals will leave the study long before they were expected to, having had little exposure to the treatment. This increases the within-group variability in treatment exposure and also decreases the exposure mean for the group overall. All these processes reduce statistical power. Standard practice now calls for describing the variability in treatment exposure, determining its causes, and then relating the variability to changes in the dependent variable (Cook et al 1993). Describing the quality of treatment implementation has now become central; even explaining this variability seems to be growing in importance.

Treatment-correlated respondent attrition is particularly problematic because it leads to selection bias—a problem exacerbated in social experiments because the treatments being contrasted often differ in intrinsic desirability and attrition is generally higher the less desirable the intervention. Field experimenters have now accumulated a set of practices that reduce the severity of treatment-correlated attrition, though none is perfect. Taken together, they are often effective. They include making generous payments to controls for providing outcome data; monitoring attrition rates early to describe them, to elicit their possible causes, and to take remedial action; following up drop-outs in whatever archival databases relevant information is located; and designing the experiment to contrast different treatments that might achieve the desired outcome as opposed to contrasting a single treatment group with a no-treatment control group. The presumption is that rival contenders as effective interventions are likely to be more similar in intrinsic desirability than are a treatment and a no-treatment control group.

Social experimenters cannot take for granted that a perfect random assignment plan will be perfectly implemented. Social experiments have to be closely monitored, even if this means adding to the already complex set of elements that might explain why a treatment impacts on an outcome. Note the two purposes of monitoring: to check on the quality of implementation of the research design, including random assignment; and to check on the quality of implementation of the intervention program itself. Variability in program implementation quality is serious, but probably not as serious as the confounding that arises when components of one treatment are shared by another. Occurring in complex field contexts, social experiments rarely permit experimental isolation and so treatments can diffuse in response to many quite different types of forces. For example, cultural mores change with the result that in one disease prevention study, members of the control group began to exercise more, eat better, and avoid drugs in response to national trends (RV Luepker et al, submitted for publication). Alternatively, no-treatment controls can and do seek alternative sources of help; educational professionals get together and share experiences so that principals from a control school may learn what is happening in an intervention school; and participants in a job training program might tell program personnel about details of the job training their friends are experiencing in other programs. Whatever the impetus, treatment diffusion is not rare, so experimenters have to look for opportunities to study groups that cannot communicate with each other. But even this does not avoid the problem that social experiments are probably most likely to be funded when a social issue is already on national policy agendas so that other solutions are simultaneously being generated in the public or private sectors. Monitoring what happens in both treatment and control groups is a *sine qua non* of modern social experiments.

## QUASI-EXPERIMENTATION

In quasi-experimentation, the slogan is that it is better to rule out validity threats by design than by statistical adjustment. The basic design features for doing this—pretests, pretest time series, nonequivalent control groups, matching, etc—were outlined by Campbell & Stanley (1963) and added to by Cook & Campbell (1979). Thinking about quasi-experiments has since evolved along three lines: 1. toward better understanding of designs that make point-specific predictions; 2. toward predictions about the multiple implications of a given causal hypothesis; and 3. toward improved analysis of data from quasi-experiments. These more recent developments are general, whereas the alternative interpretations that must be ruled out in any cause-probing study are particular to the given study (Cook & Campbell 1986). In quasi-experimental work, the local setting must be carefully examined to determine which validity threats are plausible and to identify any such threats that might be operating. With this caution in mind, we now discuss the three more recent developments mentioned above.

### *Designs Emphasizing Point-Specific Causal Hypotheses*

In interrupted time series, the same outcome variable is examined over many time points. If the cause-effect link is quick acting or has a known causal delay, then an effective treatment should lead to change in the level, slope, or variance of the time series at the point where treatment occurred. The test, then, is whether the obtained data show the change in the series at the prespecified point. Statistical conclusion validity is a potential problem because errors in a time series are likely to be autocorrelated, biasing ordinary least-squares estimates of the standard error and hence statistical tests. But internal validity is the major problem, especially because of history (e.g. some other outcome-causing event occurring at the same time as the treatment) and instrumentation (e.g. a change in record keeping occurring with the treatment). Fortunately, the point specificity of prediction limits the viability of most history and instrumentation threats to those occurring only when the treatment began. Such threats are often easily checked, are rarer than threats occurring elsewhere during the series, and can sometimes be ruled out on a priori grounds.

Plausible threats are best ruled out by using additional time series. Especially important are (*a*) control group series not expected to show the hypothesized discontinuity in level, slope, or variability of an outcome; and (*b*) additional treatment series to which the same treatment is applied at different times so we expect the obtained data to recreate the known differences in when the treatment was made available. During his career, Campbell has provided many good examples of the use of control time series and multiple time series with different treatment onset times (e.g. Campbell 1976, 1984, and other papers reproduced in Overman 1988). The design features also help when a time series is abbreviated, lacking the roughly 50 pretest and 50 posttest data points that characterize formal time series analysis. Extensions using from 5 to 20 pretest and posttest assessments are common in single-subject designs (Barlow & Hersen 1984, Kratochwill & Levin 1992).

Despite their strengths for causal hypothesis-testing, interrupted time series designs are infrequently used. Most use data already in archives because it is often costly to gather original data over long time periods. But relevant dependent variables are not always available in archives. Moreover, the relevant statistical analyses are not familiar to many social scientists, with ARIMA modeling (Box & Jenkins 1970) and spectral analysis (Granger & Newbold 1977) being particularly foreign to psychology's ANOVA tradition. But where the necessary archives exist, researchers have quickly learned these methods; and analytic strategies continue to develop (Harrop & Velicer 1985, 1990). In time series research, causal inference depends on effects occurring shortly after treatment implementation or with a known causal lag. When a possible effect is delayed, causal inference is less clear—as when job training affects income years later. Causal inference is also less clear when an intervention slowly diffuses through a population (e.g. when an AIDS prevention program is first provided to a small proportion of local needle-sharing drug users and then slowly extends to others as they are found and program resources grow). With unpredicted causal lags, time series have an ambiguous relationship to the point of treatment implementation—a problem that is compounded whenever treatment implementation is poor or highly variable. Nonetheless, interrupted time series rightly enjoy a special status among quasi-experimental designs wherever they are feasible.

Like interrupted time series studies, regression discontinuity studies also rely on the hypothesis that observations will depart from an established pattern at a specified point on a continuum. In this case the continuum is not time, but rather the regression line characterizing the relationship between outcome and an assignment variable. The regression discontinuity design depends on units on one side of an eligibility cutoff point being assigned one treatment, while units on the other side are not so assigned (e.g. income defines eligibility for Medicaid and grades determine who makes the Dean's List). More complex assignment strategies are possible using multiple cutoffs or multiple assignment variables (Trochim 1984). But in all cases, the assignment variables must be observed (not latent), and adherence to the cutoff must be strict.

The regression discontinuity design is so named because a regression line is plotted to relate the assignment and outcome variables. If the treatment is effective, a discontinuity in the regression line should occur at the cutoff point. Individuals whose income is just above and just below the eligibility point for Medicaid should differ in health status if the program is effective. The point specificity of such a prediction makes regression discontinuity resemble time series, but it is also like a randomized experiment in that the assignment of subjects to conditions is completely known. As a result, Mosteller (1990, p. 225) defines regression discontinuity as a true experiment and Goldberger (1972a,b) and Rubin (1977) have provided formal statistical proofs that regression discontinuity provides an unbiased estimate of treatment effects, just like the randomized experiment.

The widespread endorsement of the regression discontinuity design and its frequent reinvention in different disciplines suggests its usefulness whenever merit or need determine treatment assignment. But regression discontinuity studies are unfortunately even rarer than interrupted time series studies. This is partly because assignment is not always done according to strict public criteria. Professional judgment and cronyism play some role, as does the use of multiple criteria, some of which are judgmental. The low use may also be because the analysis requires accurate modeling of the functional form of the assignment-outcome relationship. Researchers typically plot linear relationships between the two variables, but if the underlying relationship is curvilinear, such modeling will yield inaccurate estimates of the discontinuity at the cutoff. Since the randomized experiment is not subject to this problem, Trochim & Cappelleri (1992) argue for combining regression discontinuity with randomized assignment, using the latter to verify the functional form at the points most open to doubt. If such a combination is not feasible, then ad hoc analyses may be needed to model functional form (Trochim 1984) or statistical tests can be used that do not make strong assumptions about such form (Robbins & Zhang 1988, 1989). However, these last tests are still in their infancy and should be treated with caution. Finally, regression discontinuity is less used because it requires 2.5 times as many subjects to achieve as much power as randomized experiments that use the pretest as a covariate, although in the absence of the pretest covariate, the power of the two designs is similar (Goldberger 1972a). In practice, power is further eroded in regression-discontinuity designs as (*a*) as the cutoff point departs from the assignment variable mean; (*b*) cases are assigned in violation of the cutoff, as with the fuzzy cutoff design discussed by Trochim (1984); or (*c*) the units assigned to a treatment fail to receive it. While these problems are also found with randomized experiments, they are likely to have more serious practical consequences for regression-discontinuity designs because researchers using such designs usually have less control over the treatment assignment implementation processes. Experimenters are therefore correct to prefer random assignment over regression-discontinuity.

The promise of the regression-discontinuity design is that it can be used whenever policy dictates that special need or merit should be a prerequisite for access to the particular services whose effectiveness is to be evaluated. In this circumstance, the treatment and control group means will almost certainly differ at the pretest, seeming to create a fatal selection confound. But this is not the case with regression-discontinuity designs because the effect estimate is not the difference between raw posttest means. It is the size of the projected discontinuity at the cutoff, which is unaffected by the correlation between the assignment and outcome variables. The design is also flexible in that many variables other than quantified need or merit can be used for treatment assignment, including the order with which individuals apply to be in a social program or the year they were born (Cain 1975). Despite the design's flexibility and impeccable logic, many practicing researchers are still skeptical. It seems implausible to them that a cutoff-based assignment process that necessarily creates a selection difference can rule out selection! Yet this is the case. Realizing that the major impediment to greater use of this design are issues of its implementability and persuasiveness in ruling out selection, Trochim & Cappelleri (1992) recently have described several variations of the design that are particularly useful for practitioners and that create the groundwork for increased use of this design in the medical context from which their examples come.

## *Designs Emphasizing Multivariate-Complex Causal Predictions*

Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations. The design elements to be combined for such prediction include (*a*) nonequivalent dependent variables, only a subset of which is theoretically responsive to a treatment, though all the dependent variables are responsive to the other plausible alternative explanations of an outcome change; (*b*) designs where a treatment is introduced, removed, and reintroduced to the same group; (*c*) nonequivalent group designs that have two or more pretest measurement waves providing a pre-intervention estimate of differences in rates of change between nonequivalent groups; (*d*) nonequivalent group designs with multiple comparison groups, some of which initially outperform the treatment group and some of which underperform it (Holland 1986); (*e*) cohort designs that use naturally occurring cycles in families or educational institutions to construct control groups of siblings or next year's freshmen that are initially less different than controls constructed in almost any other way; (*f*) other designs that match units on demonstrably stable attributes to reduce initial group nonequivalence without causing statistical regression (Holland 1986); and (*g*) designs that partition respondents or settings—even after the fact—to create subgroups that differ in treatment exposure levels and so in expected effect size.

These design features are often discussed separately. But the better strategy is to combine many of them in a single research project so as to increase the number and specificity of the testable implications of a causal hypothesis. Cook & Campbell (1979) give two examples. One example is an interrupted time series (*a*) to which a no-treatment control series was added; (*b*) where the intervention was later given to controls; and (*c*) where the original treatment series was partitioned into two nonequivalent series, only one of which the treatment should theoretically have influenced. A second example is a regression-discontinuity study of Medicaid's effects on physician visits. Medicaid eligibility depends on household income and family size. The regression-discontinuity design related income (the assignment variable) to the number of physician visits after Medicaid was passed, and it supplemented this with data on income and physician visits from the prior year. A discontinuity in the number of physician visits occurred at the cutoff point in the year after Medicaid was introduced but not in the year before.

Meehl (1978) has lamented the social sciences' dependence on null hypothesis testing, arguing that with large enough samples any null hypothesis can be rejected under some conditions. He counsels testing exact numerical predictions. Among social scientists, his advice has largely fallen on deaf ears, presumably because, outside of some areas within economics, so few theories are specific enough to make point predictions. The quasi-experimental emphasis on point-specific and multivariate-complex causal predictions approaches the spirit of Meehl's counsel, but substitutes specificity of time point predictions and pattern matching for numeric point predictions. Both point-specific and multivariate-complex predictions preserve the link to falsification because the pattern of obtained relationships facilitates causal inference only to the extent no other theory makes the same prediction about the pattern of the outcome data. Point-specific and multivariate-complex predictions lower the chance that plausible alternatives will make exactly the same prediction as the causal hypothesis under test; but they cannot guarantee this.

### *Statistical Analysis of Data from the Basic Nonequivalent Control Group Design*

Despite the growing advocacy of designs making point-specific or multivariate complex predictions (Cook 1991a), the most frequently employed quasi-experiment still involves only two (nonequivalent) groups and two measurement waves, one a pretest and the other a posttest measured on the same instrument. This design is superficially like the simplest randomized experiment and is easy to implement, perhaps explaining its great popularity. However, causal inferences are not easy to justify from this design.

Interpretation depends on identifying and ruling out all the validity threats that are plausible in the specific local context where a treatment has been implemented. But many researchers fail to consider context-specific threats, relying instead on Campbell & Stanley's (1963) checklist of validity threats when this design is used, though such threats are general and may fail to reflect unique local threats. Unfortunately, causal inferences from the design have proven far more vulnerable to model misspecification than was originally hoped. In principle, a well-specified model requires complete knowledge and perfect measurement either (*a*) of the selection process by which respondents end up in different treatment groups; or (*b*) of all causes of the outcome that are correlated with participation in treatment. The hope was that such knowledge might become available or that approximations would allow us to get quite close to the correct answer. But few social researchers who study the matter closely now believe they can identify the completely specified model, and the 1980s and early 1990s were characterized by growing disillusionment about developing even adequate approximations. Nearly any observed effect might be either nullified or even reversed in the population, depending, for example, on the variables omitted from the model (whose importance one could never know for certain) or on the pattern of unreliability of measurement across the variables in the analysis. As a result, it proved difficult to anticipate with certainty the direction of bias resulting from the simpler types of nonequivalent control group studies. But there has been much work on data analysis for this design (see Moffitt 1991, Campbell 1993).

SUGGESTIONS ABOUT DATA ANALYSIS FROM THE QUASI-EXPERIMENTAL TRADITION Most of the scholars writing about the analysis of quasi-experimental data from within the Campbell tradition are reluctant apologists for their work, preferring randomized experiments but realizing that they are not always possible. Moreover, even if such experiments are implemented initially, treatment-correlated attrition often occurs, leaving the data to be analyzed as though from a quasi-experiment. Initially, the obvious solution seemed to be to measure pretest differences between the groups under contrast and then to use simple analysis of covariance to adjust away these differences (Reichardt 1979). But two problems arise here. First, unreliability in the covariates leads to biased estimates of treatment effect (Lord 1960) so that a preference emerged for using latent variable models in which multiple observed measures of each construct are analyzed, thereby reducing measurement error and permitting an analysis only of the shared variance (Magidson 1977, Magidson & Sorbom 1982; but for a dissenting voice, see Cohen et al 1990). However, this does not deal with the second and more serious problem—that of validly specifying either the selection process or the outcome model. For this, the recommendation has been that investigators must rely on their best common sense and the available research literature in the hope of including in the analysis all likely predictors of group membership (the selection model) or outcome (the outcome model), or both (although either by itself is sufficient). Since the latent variable tradition offers no guarantee that all group differences have been statistically adjusted for, members of the theory group around Campbell turned instead to strengthening causal inference by design rather than statistical analysis (e.g. Cook 1991a) or by conducting multiple data analyses under different plausible assumptions about the direction and nature of selection bias instead of relying on a single analysis (Reichardt & Gollub 1986, Rindskopf 1986, Shadish et al 1986). The one perfect analysis remained an impossible dream, given only two nonequivalent treatment groups and two measurement waves.

SUGGESTIONS FROM LABOR ECONOMISTS At the same time, some labor economists developed an alternative tradition that sought to create statistical models that do not require full model specification and that would apply to all nonexperimental research where respondents are divided into levels on some independent variable. The best developed of this class of analyses is the instrumental variable approach associated with Heckman (Heckman 1980, Heckman et al 1987, Heckman & Hotz 1989). The probability of group membership is first modeled in a selection equation, with the predictors being a subset of variables that are presumably related to selection into the different treatment conditions. When this equation correctly predicts group membership, it is assumed that a correct selection model has been identified. The results of this model are then used as an instrumental variable in the main analysis to estimate treatment effects by adjusting for the effects of selection. Conceptually and analytically, this approach closely resembles the regression discontinuity design because it is based on full knowledge and perfect measurement of the selection model. However, in approaches like Heckman's, the selection model is not fully known. It is estimated on the basis of fallible observed measures that serve to locate more general constructs. This limitation may explain why empirical probes of Heckman's theory have not produced confirmatory results. The most compelling critiques come from studies where treatment-effect estimates derived from Heckman's selection modeling procedures are directly compared to those from randomized experiments—presumed to be the gold standard against which the success of statistical adjustment procedures should be assessed. In a reanalysis of annual earnings data from a randomized experiment on job training, both LaLonde (1986, LaLonde & Maynard 1987) and Fraker & Maynard (1987) have shown that (*a*) econometric adjustments from Heckman's work provide many different estimates of the training program's effects; and (*b*) none of these estimates closely coincides with the estimate from randomized controls. These reanalyses have led some labor economists (e.g. Ashenfelter & Card 1985, Betsey et al 1985) to suggest that unbiased estimates can only be justified from randomized experiments, thereby undermining Heckman's decade-long work on adjustments for selection bias. Other labor economists do not go quite so far and prefer what they call natural experiments—what Campbell calls quasi-experiments—over the non-experimental data labor economists previously used so readily (e.g. Card 1990, Katz & Krueger 1992, Meyer & Katz 1990; and see especially Meyer 1990 for an example that reinvents regression-discontinuity).

In Heckman & Hotz's (1989, Heckman et al 1987) rejoinder to their critics, they used the same job training data as LaLonde to argue that a particular selection model—based on two separate pretreatment measures of annual income—met certain specification tests and generated an average causal effect no different from the estimate provided by the randomized experiment. Unfortunately, their demonstration is not very convincing (Cook 1991a, Coyle et al

1991). First, there is a problem of model generality since the two-wave selection process that fit for youths who were just entering the labor market did not fit for mothers enrolled in the Aid to Families with Dependent Children Program (AFDC) who were returning to the job market. Heckman invoked a simpler cross-sectional model to describe the selection process for AFDC mothers, but this model could not be subjected to restriction tests and was assumed to be correct by fiat. Second, close inspection of the data Heckman presented reveals that the econometric analyses yielded such large standard errors that finding no difference between the two methods reflects a lack of statistical power in the econometric approach. Finally, the procedure that Heckman & Hotz found to be better for their one population had been widely advocated for at least a decade as a superior quasi-experimental design because the two pretest waves allow estimation of the preintervention selection-maturation differences between groups (Cook & Campbell 1979). Moffitt's (1991) rediscovery of the double-pretest design should be seen in the same historical light that exemplifies the quasi-experimental slogan: better to control through design than measurement and statistical adjustment. To conclude, empirical results indicate that selection models usually produce different effect estimates than do randomized experiments, and if researchers using such models were close to the mark in a particular project, they would not know this unless there were also a randomized experiment on the same topic or some better quasi-experiment. But then the need for less interpretable selection modeling would not exist!

SUGGESTIONS FROM MATHEMATICAL STATISTICIANS Historically, statisticians have preferred not to deal with the problems arising from quasi-experiments (which they call observational studies). They almost monolithically advocate randomized designs unless prior information is so reliable and comprehensive that Bayes' theorem is obviously relevant. This situation changed a little in the 1980s, and a flurry of publications appeared guided by the realization that quasi-experiments were rampant and might be improved upon even if never perfected (Holland 1986; Rosenbaum 1984; Rosenbaum & Rubin 1983, 1984).

The statisticians provided four major suggestions for improving the analysis of quasi-experimental data. The first emphasized the need for selecting study units matched on stable attributes, not as a substitute for random assignment, but as one of a series of palliatives that reduce bias (Rubin 1986, Rosenbaum 1984). The second was a call to compute a propensity score, an empirical estimate of the selection process based, to the extent possible, on actual observation of which units enter the various treatment groups being contrasted (Rosenbaum & Rubin 1983, 1984). The intent is to generate the most accurate and complete description of the selection process possible rather than to find a proxy for the entire selection process. The third improvement the statisticians counseled was the use of multiple comparison groups to increase statistical power; to rule out construct validity threats through the use of placebo controls, for example; and to vary the direction of pretest group differences and hence the assumptions made about growth rate differences between groups. The statisticians want to avoid having a single comparison group that outperforms or underperforms a treatment group before the treatment and might be changing over time at a different rate from the treatment group. The final suggestion was that multiple data analyses should be conducted under different explicit and plausible assumptions about factors that might influence posttest performance in one group more than another. Clearly, the statisticians see quasi-experimental analysis as involving a more complicated, thoughtful, theory-dependent process of argument construction than is the case when analyzing data from randomized experiments.

It is interesting to note some convergences across all the different traditions of dealing with group nonequivalence. One is the advocacy of directly observing the selection process so as to have a more fully specified and more accurate model. Confidence in instrumental variables or single pretest measures as adequate proxies is waning. A second convergence is the advocacy of multiple control groups, especially to make heterogeneous the direction of pretest bias. A third is the advisability of using several pretreatment measurement waves to observe how the different groups might be changing spontaneously over time. A fourth is the value of using matching to minimize initial group noncomparability provided that the matching will not lead to statistical regression. The final point of convergence is that multiple data analyses should take place under very different (and explicit) assumptions about the direction of bias. The aim is to test the robustness of results across different plausible assumption sets. We see here the call for an honest critical multiplism (Cook 1985, Shadish 1989); a lack of confidence in the basic two-group, two-wave quasi-experimental design that is so often used in social experiments; and the need for more specific causal hypotheses, supplementing the basic design with more measurement waves and more control groups.

During the years we are examining, the use of empirical research on experiments has been increasing. From the beginning, Campbell (1957) argued that his lists of validity threats could be shortened or lengthened as experience accumulated about the viability of particular validity threats or the necessity to postulate new ones. Today we see even more focused research on design issues, as with the labor economists who contrast the results from randomized experiments and quasi-experiments. The use of empirical research to improve design is also evident in meta-analysis, particularly as it explores how methodological decisions influence effect size estimates. For example, do published studies yield different estimates from unpublished studies (Shadish et al 1989)? Does the presence of pretest measurement change effect size estimates

(Willson & Putnam 1982)? Do randomized experiments yield different estimates than quasi-experiments in general or than certain types of quasi-experiments in particular (Heinsman 1993)? We expect explicit meta-analytic studies of design questions to become even more prevalent in the future.

## GENERALIZING CAUSAL RELATIONSHIPS

Cronbach's (1982) *UTOS* formulation suggests that experimenters often hope that the causal inferences they claim (*a*) will generalize to specific populations of units and settings and to target cause and effect constructs; and (*b*) can be extrapolated to times, persons, settings, causes, and effects that are manifestly different from those sampled in the existing research. Campbell's (1957) work on external validity describes a similar but less articulated aspiration. But such generalization is not easy. For compelling logistical reasons, nearly all experiments are conducted with local samples of convenience in places of convenience and involve only one or two operationalizations of a treatment, though there may be more operationalizations of an outcome. How are Cronbach's two types of generalization to be promoted if most social experiments are so contextually specific?

### *Random Sampling*

The classic answer to this question requires randomly sampling units from the population of persons, settings, times, treatments, or outcomes to which generalization is desired. The resulting samples provide unbiased estimates of the population parameters within known sampling limits, thereby solving Cronbach's first causal generalization problem. Unfortunately, this solution is rarely applicable. In single experiments it is sometimes possible to draw a formal probability sample of units and even settings. However, it is almost impossible to draw probability samples of treatments or outcomes because no enumeration of the population can be achieved given that the treatment and outcome classes can seldom be fully defined. Occasionally, an enumeration of settings or units is possible, as with studies intending to generalize to elementary school children in a particular city, or to all community mental health centers in the nation. But for dispersed populations, there are enormous practical difficulties with drawing the appropriate sample in such a way that the treatment can be implemented with high quality. Even if these difficulties could be surmounted, the experiment would usually still be restricted to those individuals who agreed to be randomly assigned—surely only a subpopulation of all those to whom generalization is sought. Moreover, attrition is likely after an experiment has begun, making the achieved population even less similar to the target population. For all these reasons, formal probability sampling methods seem unlikely to provide a general solution to Cronbach's first framing of the causal generalization problem, though they are desirable whenever feasible.

## *Proximal Similarity*

Campbell's solution to Cronbach's first generalization problem is implicit in his tentative relabeling of external validity as proximal similarity, by which he means that "as scientists we generalize with most confidence to applications most similar to the setting of the original research" (Campbell 1986, p. 75). He believes that the more similar an experiment's observables are to the setting to which generalization is desired, the less likely it is that background variables in the setting of desired application will modify the cause-effect relationship obtained in the original experiment. Cronbach appears to agree with this. But applying this rationale requires knowledge of which attributes must be similar, and this presumably depends on good theory, validated clinical experience, or prior relevant experimental results. Alas, good substantive theory is not always available, intuitive judgments can be fickle, and prior findings are often lacking. What can be done in such cases?

St. Pierre & Cook (1984) suggest two related purposive sampling options that are useful in planning generalizable cause-probing experiments. In modal instance sampling, the researcher uses background information or pilot work to select experimental particulars that proximally resemble the kinds of unit, setting, treatment, and outcome variants that are most commonly found in the settings to which generalization is sought. The strategy here is to capture the most representative instances (Brunswick 1955). In sampling to maximize heterogeneity, the aim is to sample as widely as possible along dimensions that theory, practice, or other forms of speculation suggest might codetermine the efficacy of a treatment. If a treatment is robustly effective at widely differing points on such dimensions, then generalization over them is facilitated. If the treatment is not general in its effects, then the researcher can conduct subgroup analyses to identify some of the specific boundary conditions under which the treatment is effective. Since power is potentially a problem when examining statistical interactions, St. Pierre & Cook (1984) also suggest probing whether a cause-effect relationship can be demonstrated across whatever heterogeneity is deliberately allowed into the sampling plan. Purposive sampling either to increase heterogeneity or to capture modal instances does not justify generalization as well as random sampling does, but it should increase insights into how broadly a causal relationship can be generalized.

## *Probes for Robust Replication over Multiple Experiments*

A collection of experiments on the same or related topics can allow even better probes of the causal contingencies influencing treatment effectiveness. This is because each experiment is likely to involve a different population of persons and settings, a different time, and unique ways of conceptualizing and measuring the cause and effect. A search can then take place to identify which causal relationships are robust across the obtained variation in all these dimensions and also to identify those contexts that limit generalization because no cause-effect relationship can be found in them. Meta-analysis is the best-known exemplar of this approach to causal generalization (Cook 1991b, Cook et al 1992).

Consider meta-analytic work on the effects of psychotherapy. This suggests that effect sizes do not differ across such seemingly crucial dimensions as therapist training and experience, treatment orientation, and duration of treatment (Smith et al 1980, Berman & Norton 1985). However, effect sizes do differ by outcome characteristics, with larger effects emerging for outcomes that receive the most attention in treatment (Shadish et al 1993), suggesting one restriction to the generality of psychotherapy effects. Empirical efforts to probe the robustness of a cause-effect relationship require the pool of available studies to vary on factors likely to moderate the relationship. But sometimes there is little or no variability. For example, in psychotherapy literature there are very few experimental studies of psychoanalytic orientations, so we still do not know if the general finding about therapist orientation making little difference to client outcomes also applies to psychoanalysis as a particular form of treatment.

Still, with the greater range and heterogeneity of sampling that literature reviews promote it is often possible to probe whether a causal relationship is reproduced in contexts that are proximally similar to particular, defined contexts of intended application. This speaks to Cronbach's first framing of causal generalization. Important for his second framing is that when a causal relationship demonstrably holds over a wide array of different contexts (e.g. Devine 1992), it seems reasonable to presume that it will continue to hold in future contexts that are different from past ones. The warrant for such extrapolation is logically flawed; nonetheless, it is pragmatically superior to the warrant for extrapolation when a relationship has only been tested within a narrow range of persons, settings, times, causal manipulations, and effect measures.

## *Causal Explanation*

We cannot expect all relevant causal contingencies to be represented in the body of studies available for synthesis. Hence, we need other methods for extrapolating a causal connection to unstudied populations and classes. Cronbach believes that causal explanation is the key to such generalization, and that causal explanation requires identifying the processes that follow because a treatment has varied and without which the effect would not have been produced. Identifying such micro-mediating processes usually involves decomposing the molar cause and effect constructs into the subsets of components thought to be most critically involved in the molar causal link and then examining all micro-mediating processes that might link the causally efficacious treatment components to the causally impacted outcome components.

The presumption is that once such knowledge has been gained, the crucial causal processes can be transferred to novel contexts of application. Cook & Campbell (1979) use the example of a light switch to illustrate this. Knowledge that flicking the switch results in light is the type of descriptive knowledge about manipulanda that experiments promote; more explanatory knowledge requires knowing about switch mechanisms, wiring and circuitry, the nature of electricity, and how all these elements can be combined to produce light. Knowing so much increases the chances of creating light in circumstances where there are no light switches, providing one can reproduce the causal explanatory processes that make light in whatever ways local resources allow. One need not be restricted to what can be bought in a hardware store or to what was available when one first learned about electricity. Another example concerns the murder of Kitty Genovese, who was stabbed repeatedly for over half an hour, with nearly forty neighbors watching and not offering to help. Latane & Darley (1970) hypothesized that this apathy resulted from a process they called the diffusion of responsibility—everyone thought someone else would surely help. Latane & Darley experimentally demonstrated that this process was common to a host of other situations as diverse as when a woman falls and sprains her ankle, smoke comes into a room from a fire, someone has an epileptic seizure, or a cash register is robbed (Brown 1986). Once micro-mediating processes are identified, they are often broadly transferable, certainly more so than if the research had shown that helping is reduced only following a stabbing and then only in New Jersey and only with apartment dwellers.

Identifying causal explanatory processes is not easy. Randomized experiments were designed to provide information about descriptive causal connections and not about processes accounting for these connections. Hence, the methods for exploring causal mediation must be added to experimental frameworks. Within quantitative research traditions, potential mediators can be assessed and used in some form of causal modeling. Alternatively, respondents might be assigned to treatments that vary in the availability of a hypothesized explanatory mechanism. However, experience with complex randomized experiments has not been salutary in field research. The Negative Income Tax experiments are among the most complex ever designed with respect to the number of factors examined, the number levels on each factor, and the unbalanced distribution of respondents across factors and levels (Kershaw & Fair 1976). Preserving all these distinctions in field settings has proven quite difficult, though some fairly complex planned variation studies have been carried out successfully (Connell et al 1985). Cronbach (1982) has proposed that the qualitative methods of the ethnographer, historian, or journalist should also be used to generate and defend hypotheses about micro-mediating processes. Although we personally have a lot of sympathy for this qualitative approach, it is likely to fall on deaf ears in the social science community at large.

## CONCLUSIONS

Field experiments are now much more commonplace than they were twenty years ago. The pendulum has swung to favor randomized experiments over quasi-experiments—and strongly so in all areas except perhaps intervention research in education. Considerable informal information is now available about the many factors that facilitate the implementation of random assignment and about factors that promote better randomized studies. As a result, fields like labor economics and community health promotion that fifteen years ago were characterized by the use of statistical adjustments to facilitate causal inference now use more (and larger) experiments and fewer quasi-experiments.

Though quasi-experiments have lost some of their warrant, some progress has been made toward improving them. More attention has been directed to two designs that promote point-specific causal inferences—the regression-discontinuity and interrupted time series designs (Marcantonio & Cook 1994). Also more often advocated are somewhat elaborate quasi-experimental designs that predict a complex pattern of results emanating from the causal agent under analysis. At issue here are more pretest measurement waves, more control groups, better matched controls (including cohorts), and removing and reinstating the treatment at known time points (Cook et al 1990).

Much work has also gone into the statistical analysis of data from quasi-experiments in order to model selection biases. There are several different traditions that, unfortunately, rarely communicate with each other (see Campbell 1993). This is sad because they are growing ever closer together. None of these traditions is sanguine about finding the elusive demonstrably valid counterfactual baseline unless the selection model is completely known, as with randomized experiments or the regression-discontinuity design. All traditions agree, though, that causal inference is better warranted the more closely the comparison groups are matched, the more information there is to describe the selection process, the more comparison groups there are (particularly if they bracket the treatment group mean at the pretest), and the more robust are the results generated from multiple statistical analyses made under different plausible assumptions about the nature of selection bias.

But selection is not the only inferential problem researchers need to worry about. Campbell's (1957) internal validity threats of testing, instrumentation, and history are often relevant, as are Cook & Campbell's (1979) threats of treatment diffusion, resentful demoralization, compensatory rivalry, and compensatory equalization. Statistical research on selection modeling has been dominant in the past, and rightly so. But it should not crowd out research into these other relevant threats to valid causal inference.

Work on the generalization of causal relationships has historically been less prevalent than work on establishing causal relationships. There is ongoing work on the generalization of causal inferences in two modes. One is meta-analytic and emphasizes the empirical robustness of a particular causal connection across a wide range of persons, settings, times, and cause and effect constructs. The second is more causal-explanatory, and it emphasizes identifying the micro-mediating processes that causally connect a treatment to an outcome, usually through a process of theoretical specification, measurement, and data analysis rather than through the sampling strategy that characterizes meta-analysis. Causal generalization is now more of an issue in social experiments than it was fifteen years ago (Cook 1991a, 1993; Friedlander & Guerin 1992).

## *Literature Cited*

Aiken LS, West SG. 1991. *Multiple Regression: Testing and Interpreting Interactions.* Newbury Park, CA: Sage

Ashenfelter O, Card D. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. *Rev. Econ. Stat.* 67:648-60

Barlow DH, Hersen M, eds. 1984. *Single Case Experimental Designs: Strategies for Studying Behavior Change.* New York: Pergamon. 2nd ed.

Berman JS, Norton NC. 1985. Does professional training make a therapist more effective? *Psychol. Bull.* 98:401–7

Betsey CL, Hollister RE, Papageorgiou MR, eds. 1985. *Youth Employed and Training Programs: The YEDPA Years.* Washington, DC: Natl. Acad. Press

Bickman L. 1987. Functions of program theory. In *Using Program Theory in Evaluation: New Directions for Program Evaluation,* ed. L Bickman, 33:5–18. San Francisco: Jossey-Bass

Bickman L, ed. 1990. *Advances in Program Theory: New Directions for Program Evaluation,* Vol. 47. San Francisco: Jossey-Bass

Boruch RF. 1975. On common contentions about randomized experiments. In *Experimental Tests of Public Policy,* ed. RF Boruch, HW Riecken, pp. 107–42. Boulder, CO: Westview

Boruch RF, Gomez H. 1977. Sensitivity, bias and theory in impact evaluations. *Prof. Psychol. Res. Pract.* 8:411–34

Boruch RF, McSweeney AJ, Soderstrom EJ. 1978. Randomized field experiments for program planning development and evaluation. *Eval. Q.* 2:655–95

Boruch RF, Wothke W, eds. 1985. *Randomization and Field Experimentation: New Directions for Program Evaluation,* Vol. 28. San Francisco: Jossey-Bass

Box GEP, Jenkins GM. 1970. *Time-series Analysis: Forecasting and Control.* San Francisco: Holden-Day

Brown R. 1986. *Social Psychology: The Second Edition.* New York: Free Press

Brunswik E. 1955. *Perception and the Representative Design of Psychological Experiments.* Berkeley: Univ. Calif. Press. 2nd ed.

Bryk AS, Raudenbush SW. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage

Cain GG. 1975. Regression and selection models to improve nonexperimental comparisons. In *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs,* ed. CA Bennett, AA Lumsdaine, pp. 297–317. New York: Academic

Campbell DT. 1957. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* 54:297–312

Campbell DT. 1966. Pattern matching as an essential in distal knowing. In *The Psychology of Egon Brunswik,* ed. KR Hammond, pp. 81–106. New York: Holt, Rinehart, Winston

Campbell DT. 1976. Focal local indicators for social program evaluation. *Soc. Indic. Res.* 3:237–56

Campbell DT. 1982. Experiments as arguments. *Knowl.: Creation Diffus. Util.* 3:237–56

Campbell DT. 1984. Hospital and landsting as continuously monitoring social programs: advocacy and warning. In *Evaluation of Mental Health Service Programs,* ed. B Cronhom, L von Korring, pp. 13–39. Stockholm: Forskningsraadet Medicinska

Campbell DT. 1986. Relabeling internal and external validity for applied social scientists. See Trochim 1986, pp. 67–77

Campbell DT. 1993. Quasi-experimental research designs in compensatory education. In *Evaluating Intervention Strategies for Children and Youth at Risk,* ed. EM Scott. Washington, DC: US Govt. Print. Office. In press

Campbell DT, Boruch RF. 1975. Making the case for randomized assignment to treatments by considering the alternatives: six ways in which quasi-experimental evaluations tend to underestimate effects. In *Evaluation and Experience: Some Critical Issues in Assessing Social Programs,* ed. CA Bennett, AA Lumsdaine, pp. 195–296. New York: Academic

Campbell DT, Stanley JC. 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand-McNally

Card D. 1990. The impact of the Mariel boatlift on the Miami labor market. *Ind. Labor Relat.* 43:245–57

Chelimsky E. 1987. The politics of program evaluation. *Soc. Sci. Mod. Soc.* 25(1):24–32

Chen HT, Rossi PH. 1983. Evaluating with sense: the theory-driven approach. *Eval. Rev.* 7:283–302

Cohen P, Cohen J, Teresi J, Marchi M, Velez NC. 1990. Problems in the measurement of latent variables in structural equations causal models. *Appl. Psychol. Meas.* 14(2):183–96

Collingwood RG. 1940. *An Essay on Metaphysics.* Oxford: Clarendon

Connell DB, Turner RT, Mason EF. 1985. Summary of findings of the school health education evaluation: health promotion effectiveness, implementation, and costs. *J. School Health* 85:316–17

Cook TD. 1985. Post-positivist critical multiplism. In *Social Science and Social Policy,* ed. RL Shotland, MM Mark, pp. 21–62. Beverly Hills, CA: Sage

Cook TD. 1991a. Clarifying the warrant for generalized causal inferences in quasi-experimentation. In *Evaluation and Education at Quarter Century,* ed. MW McLaughlin, D Phillips, pp. 115–44. Chicago: NSSE

Cook TD. 1991b. Meta-analysis: its potential for causal description and causal explanation within program evaluation. In *Social Prevention and the Social Sciences: Theoretical Controversies. Research Problems and Evaluation Strategies,* ed. G Albrecht, H-U Otto, S Karstedt-Henke, K Bollert, pp. 245–85. Berlin-New York: de Gruyter

Cook TD. 1993. A quasi-sampling theory of the generalization of causal relationships. In *Understanding Causes and Generalizing About Them: New Directions for Program Evaluation,* ed. L Sechrest, AG Scott, 57:39–82. San Francisco: Jossey-Bass

Cook TD, Anson A, Walchli S. 1993. From causal description to causal explanation: improving three already good evaluations of adolescent health programs. In *Promoting the Health of Adolescents: New Directions for the Twenty-First Century,* ed. SG Millstein, AC Petersen, EO Nightingale, pp. 339–74. New York: Oxford Univ. Press

Cook TD, Campbell DT. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand-McNally

Cook TD, Campbell DT. 1986. The causal assumptions of quasi-experimental practice. *Synthese* 68:141–80

Cook TD, Campbell DT, Perrachio L. 1990. Quasiexperimentation. In *Handbook of Industrial and Organizational Psychology,* ed. MD Dunnette, LM Hough, pp. 491–576. Palo Alto, CA: Consult. Psychol. Press. 2nd ed.

Cook TD, Cooper H, Cordray D, Hartmann H, Hedges L, Light R, Louis T, Mosteller F, eds. 1992. *Meta-Analysis for Explanation: A Casebook.* New York: Russell Sage Found.

Coyle SL, Boruch RF, Turner CF, eds. 1991. *Evaluating AIDS Prevention Programs: Expanded Edition.* Washington, DC: Natl. Acad. Press

Cronbach LJ. 1982. *Designing Evaluations of Educational and Social Programs.* San Francisco: Jossey-Bass

Cronbach LJ, Snow RE. 1976. *Aptitudes and Instructional Methods.* New York: Irvington

Devine E. 1992. Effects of psychoeducational care with adult surgical patients: a theory-probing meta-analysis of intervention studies. See Cook et al 1992, pp. 35–84.

Dunn WN. 1982. Reforms as arguments. *Knowl. Creation Diffus. Util.* 3(3):293–326

Fraker T, Maynard R. 1987. Evaluating the adequacy of comparison group designs for evaluation of employment-related programs. *J. Hum. Res.* 22:194–227

Freedman DA. 1987. A rejoinder on models, metaphors, and fables. *J. Educ. Stat.* 12:206–23

Friedlander D, Gueron JM. 1992. Are high-cost services more effective than low-cost services? In *Evaluating Welfare and Training Programs,* ed. CF Manski, I Garfinkel, pp. 143–98. Cambridge: Harvard Univ. Press

Gadenne V. 1976. *Die Gultigkeit psychologischer Untersuchungen.* Stuttgart, Germany: Kohlhammer

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

Goldberger AS. 1972a. Selection bias in evaluating treatment effects: some formal illustrations. In *Discussion Papers No. 123–172.* Madison: Inst. Res. Poverty, Univ. Wisc.

Goldberger AS. 1972b. Selection bias in evaluating treatment effects: the case of interaction. In *Discussion Papers No. 123–172* Madison: Inst. Res. Poverty, Univ. Wisc.

Gottfredson SD. 1978. Evaluating psychological research reports: dimensions, reliability and correlates of quality judgments. *Am. Psychol.* 33:920–34

Granger CW, Newbold P. *Forecasting Economic Time Series.* New York: Academic

Harrop JW, Velicer WF. 1985. A comparison of alternative approaches to the analysis of interrupted time series. *Multivariate Behav. Res.* 20:27–44

Harrop JW, Velicer WF. 1990. Computer programs for interrupted time series analysis: 1. A qualitative evaluation. *Multivariate Behav. Res.* 25:219–31

Heckman JJ. 1980. Sample selection bias as a specification error. In *Evaluation Studies Review Annual,* ed. EW Stromsdorfer, G Farkas, 5:13–31. Newbury Park, CA: Sage

Heckman JJ, Hotz VJ. 1988. Are classical experiments necessary for evaluating the impact of manpower training programs? A critical assessment. *Ind. Relat. Res. Assoc. 40th Annu. Proc.,* pp. 291–302

Heckman JJ, Hotz VJ. 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *J. Am. Stat. Assoc.* 84:862–74

Heckman JJ, Hotz VJ, Dabos M. 1987. Do we need experimental data to evaluate the impact of manpower training on earnings? *Eval. Rev.* 11:395–427

Heinsman D. 1993. *Effect sizes in meta-analysis: does random assignment make a difference?* PhD thesis. Memphis State Univ., Tenn.

Holland PW. 1986. Statistics and causal inference (with discussion). *J. Am. Stat. Assoc.* 81:945–70

Holland PW, Rubin DB. 1988. Causal inference in retrospective studies. *Eval. Rev.* 12:203–31

Katz LF, Krueger AB. 1992. The effect of minimum wage on the fast-food industry. *Ind. Labor Relat.* 46:6–21

Kershaw D, Fair J. 1976. *The New Jersey Income-Maintenance Experiment.* Vol. 1: *Operations, Surveys and Administration.* New York: Academic

Kish L. 1987. *Statistical Design for Research.* New York: Wiley

Kratochwill TR, Levin JR, eds. 1992. *Single-Case Research Design and Analysis.* Hillsdale, NJ: Erlbaum

Kruglanski AW, Kroy M. 1975. Outcome validity in experimental research: a re-conceptualization. *J. Represent. Res. Soc. Psychol.* 7:168–78

Kuhn TS. 1962. *The Structure of Scientific Revolutions.* Chicago: Univ. Chicago Press

Lakoff G. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind.* Chicago: Univ. Chicago Press

LaLonde RJ. 1986. Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* 76:604–20

LaLonde RJ, Maynard R. 1987. How precise are evaluations of employment and training experiments: evidence from a field experiment. *Eval. Rev.* 11:428–51

Latane B, Darley JM. 1970. *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century-Crofts

Lipsey M. 1993. Theory as method: small theories of treatments. In *Understanding Causes and Generalizing About Them: New Directions for Program Evaluation,* ed. LB Sechrest, AG Scott, 57:5–38. San Francisco: Jossey-Bass

Lord FM. 1960. Large-scale covariance analysis when the control variable is fallible. *J. Am. Stat. Assoc.* 55:307–21

Mackie JL. 1974. *The Cement of the Universe.* Oxford: Oxford Univ. Press

Magidson J. 1977. Toward a causal model approach for adjusting for pre-existing differences in the non-equivalent control group situation. *Eval. Q.* 1(3):399–420

Magidson J, Sorbom D. 1982. Adjusting for confounding factors in quasi-experiments. *Educ. Eval. Policy Anal.* 4:321–29

Manski CF, Garfinkel I, eds. 1992. *Evaluating Welfare and Training Programs.* Cambridge: Harvard Univ. Press

Marcantonio RJ, Cook TD. 1994. Convincing quasi-experiments: the interrupted time series and regression-discontinuity designs. In *Handbook of Practical Program Evaluation,* ed. JS Wholey, HP Hatry, KE Newcomer. San Francisco: Jossey-Bass. In press

Mark MM. 1986. Validity typologies and the logic and practice of quasi-experimentation. See Trochim 1986, pp. 47–66

Medin DL. 1989. Concepts and conceptual structure. *Am. Psychol.* 44:1469–81

Meehl PE. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46:806–34

Meyer BD. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58:757–82

Meyer BD, Katz LF. 1990. The impact of the potential duration of unemployment benefits on the duration of unemployment. *J. Public Econ.* 41:45–72

Moffitt R. 1991. The use of selection modeling to evaluate AIDS interventions with observational data. *Eval. Rev.* 15(3):291–314

Mosteller F. 1990. Improving research methodology: an overview. In *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*, ed. L Sechrest, E Perrin, J Bunker, pp. 221–30. Rockville, MD: AHCPR, PHS

Overman ES. 1988. *Methodology and Epistemology for the Social Sciences: Selected Papers of Donald T. Campbell.* Chicago: Univ. Chicago Press

Popper KR. 1959. *The Logic of Scientific Discovery.* New York: Basic Books

Reichardt CS. 1979. The statistical analysis of data from nonequivalent group designs. See Cook & Campbell 1979, pp. 147–205

Reichardt CS, Gollob HF. 1986. Satisfying the constraints of causal modelling. See Trochim 1986, pp. 91–107

Riecken HW, Boruch RF, eds. 1974. *Social Experimentation.* New York: Academic

Rindskopf D. 1986. New developments in selection modeling for quasi-experimentation. See Trochim 1986

Robbins H, Zhang C-H. 1988. Estimating a treatment effect under biased sampling. *Proc. Natl. Acad. Sci. USA* 85:3670–72

Robbins H, Zhang C-H. 1989. Estimating the superiority of a drug to a placebo when all and only those patients at risk are treated with the drug. *Proc. Natl. Acad. Sci. USA* 86:3003–5

Rosch E. 1978. Principles of categorization. In *Cognition and Categorization*, ed. E Rosch, BB Lloyd. Hillsdale, NJ: Erlbaum

Rosenbaum PR. 1984. From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment. *J. Am. Stat. Assoc.* 79:41–48

Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55

Rosenbaum PR, Rubin DB. 1984. Reducing bias in observation studies for causal effects. *J. Educ. Psychol.* 66:688–701

Rubin DB. 1977. Assignment to treatment group on the basis of a covariate. *J. Educ. Stat.* 2:1–26

Rubin DB. 1986. Which ifs have causal answers? *J. Am. Stat. Assoc.* 81:961–62

Sechrest L, West SG, Phillips MA, Redner R, Yeaton W. 1979. Some neglected problems in evaluation research: strength and integrity of treatments. In *Evaluation Studies Review Annual*, ed. L Sechrest, SG West, MA Phillips, R Redner, W Yeaton, 4:15–35. Beverly Hills, CA: Sage

Shadish WR. 1989. Critical multiplism: a research strategy and its attendent tactics. In *Health Services Research Methodology: A Focus on AIDS. DHHS Pub. No. PHS89-3439*, ed. L Sechrest, H Freeman, A Mully. Rockville, MD: NCHS, USDHHS

Smith EE, Medin DL. 1981. *Categories and Concepts*, Cambridge: Harvard Univ. Press

Shadish WR, Cook TD, Houts AC. 1986. Quasi-experimentation in a critical multiplist mode. See Trochim 1986, pp. 29–46

Shadish WR, Doherty M, Montgomery LM. 1989. How many studies are in the file drawer? An estimate from the family/marital psychotherapy literature. *Clin. Psychol. Rev.* 9:589–603

Shadish WR, Montgomery LM, Wilson P, Wilson MR, Bright I, Okwumabua TM. 1993. The effects of family and marital psychotherapies: a meta-analysis. *J. Consult. Clin. Psychol.* In press

Smith EE, Medin DL. 1981. *Categories and Concepts.* Cambrigde: Harvard Univ. Press

Smith ML, Glass GV, Miller TI. 1980. *The Benefits of Psychotherapy.* Baltimore: Johns Hopkins Univ. Press

St. Pierre RG, Cook TD. 1984. Sampling strategy in the design of program evaluations. In *Evaluation Studies Review Annual*, ed. RF Conner, DG Altman, C Jackson, 9:459–84. Beverly Hills, CA: Sage

Trochim WMK. 1984. Research design for program evaluation: the regression-discontinuity approach. Newbury Park, CA: Sage

Trochim WMK, ed. 1986. *Advances in Quasi-experimental Design Analysis: New Directions for Program Evaluation*, Vol. 31. San Francisco: Jossey-Bass

Trochim WMK, Cappelleri JC. 1992. Cutoff assignment strategies for enhancing randomized clinical trials. *Control. Clin. Trials* 13:190–212

Whitbeck C. 1977. Causation in medicine: the disease entity model. *Philos. Sci.* 44:619–37

Willson VL, Putnam RR. 1982. A meta-analysis of pretest sensitization effects in experimental design. *Am. Educ. Res. J.* 19:249–58

Wortman PM. 1993. Judging research quality. In *Handbook of Research Synthesis,* ed. HM Cooper, LV Hedges. New York: Russell Sage Foundation. In press

Zimmermann HJ, Zadeh LA, Gaines BR, eds. 1984. *Fuzzy Sets and Decision Analysis.* New York: Elsevier