Annu. Rev. Psychol. 2000. 51:345–375

EVALUATION METHODS FOR SOCIAL INTERVENTION

Mark W. Lipsey and David S. Cordray
Department of Psychology and Human Development, Vanderbilt University, Nashville, Tennessee 37203; e-mail: mark.lipsey@vanderbilt.edu, david.s.cordray@vanderbilt.edu

Key Words program evaluation, treatment effectiveness, outcomes, field experiments, quasi-experiments, analysis of change

■ **Abstract** Experimental design is the method of choice for establishing whether social interventions have the intended effects on the populations they are presumed to benefit. Experience with field experiments, however, has revealed significant limitations relating chiefly to (*a*) practical problems implementing random assignment, (*b*) important uncontrolled sources of variability occurring after assignment, and (*c*) a low yield of information for explaining why certain effects were or were not found. In response, it is increasingly common for outcome evaluation to draw on some form of program theory and extend data collection to include descriptive information about program implementation, client characteristics, and patterns of change. These supplements often cannot be readily incorporated into standard experimental design, especially statistical analysis. An important advance in outcome evaluation is the recent development of statistical models that are able to represent individual-level change, correlates of change, and program effects in an integrated and informative manner.

CONTENTS

0084–6570/00/0201–0345$12.00

INTRODUCTION

The field of program evaluation presents a diversity of images and claims about the nature and role of evaluation that confounds any attempt to construct a coherent account of its methods or confidently identify important new developments. We take the view that the overarching goal of the program evaluation enterprise is to contribute to the improvement of social conditions by providing scientifically credible information and balanced judgment to legitimate social agents about the effectiveness of interventions intended to produce social benefits. Because of its centrality in this perspective, this review focuses on outcome evaluation, that is, the assessment of the effects of interventions upon the populations they are intended to benefit. The coverage of this topic is concentrated on literature published within the last decade with particular attention to the period subsequent to the related reviews by Cook and Shadish (1994) on social experiments and Sechrest & Figueredo (1993) on program evaluation.

The classical and still standard methodological paradigm for outcome evaluation is experimental design and its various quasi-experimental approximations. The superiority of experimental methods for investigating the causal effects of deliberate intervention is widely acknowledged. It is also widely acknowledged that the experimental paradigm as it is conventionally applied in outcome evaluation has significant limitations. In particular, the practice of outcome evaluation is marked by increased recognition of important variables that the experiment cannot control.

Random assignment, for instance, is recognized as a useful means of equating groups prior to the delivery of an intervention (Boruch 1997), though practical and ethical constraints often necessitate comparison of nonrandomized groups.Whereas assignment to treatment and control conditions is a defining event in outcome evaluation, decades of experience have shown that after assignment important processes occur that can seriously influence the quality of the evaluation design, the interpretability of the results, and the utility of the study. Among these processes are (*a*) poor program implementation, (*b*) augmentation of the control group with nonprogram services, (*c*) poor retention of participants in program and control conditions, (*d*) receipt of incomplete or inconsistent program services by participants, and (*e*) attrition or incomplete follow-up measurement. In addition, a host of participant characteristics (e.g. problem severity, motivation, ability level) can interact with exposure and response to treatment in ways that further complicate the situation.

Although efforts to minimize compromises to ideal experimental design should be encouraged, program circumstances typically permit little control over many variables that potentially have substantial influence on the outcomes under investigation. When such variables cannot be controlled and are too important to ignore, the only alternative is to measure them and incorporate the results into the statistical analysis. This approach not only permits some degree of statistical

control over problem variables, but also supplies more information about factors that may explain why effects were or were not found. Therefore, in outcome evaluation, data collection has been increasingly extended to include measurement of such variables as program implementation, participants' exposure to services, and those participant characteristics and responses that may mediate or moderate the effects of treatment. In order for such variables to be measured, of course, they must first be identified as relevant and be defined with sufficient specificity to be operationalized. These are conceptual tasks, and correspondingly, outcome evaluation is increasingly being guided by theories of program and individual change.

Given the nature and number of the sources of uncontrolled variability in outcome evaluation, it is hardly surprising that different patterns of change across individuals can and have emerged in evaluation studies (e.g. Collins 1996, Krause et al 1998). An important advance in outcome evaluation has been the recent development of statistical models for growth (or decay) that are able to represent individual-level change, correlates of change, and program effects in a sophisticated and informative manner.

PROBLEMS AND PROGRESS IN EXPERIMENTAL METHODS

Evaluators have identified many practical and conceptual limitations in the experimental and quasi-experimental designs that for decades have been the primary tools for investigating program effects (Campbell & Stanley 1966, Cook & Campbell 1979). Recognition of those limitations, in turn, has stimulated significant methodological innovation. Indeed, much of the history of methodological development in evaluation research can be viewed as refinements of, or reactions to, the perceived deficiencies of experimental design for certain evaluation purposes. The major areas of problems and progress, summarized below, touch on almost every aspect of experimental design in field settings.

Random Assignment

The practicality of random assignment has been a continuing issue in the study of intervention effects (Dennis 1990) despite a steady cumulation of know-how about implementing it in field settings (Boruch 1997). Although it is not hard to find examples of successful randomization (e.g. Gueron 1997, Braucht & Reichardt 1993), its applicability is not uniform. In some cases there are too few units to randomize, a notable problem for community-level programs (Murray et al 1996), or no units available to be assigned to control conditions, as with full-coverage programs. In other instances, randomization raises ethical or legal questions about withholding services from otherwise eligible persons. In response to these concerns, alternative allocation methods have been proposed, such as

sequential assignment (Staines et al 1999). However, even in favorable situations, the procedural difficulty of assigning persons to service conditions solely on the basis of chance and maintaining their participation through final data collection may thwart well-intentioned efforts. As a result, outcome evaluations often use the weakest of the quasi-experimental designs, nonequivalent comparisons, either by intent or because of degradation of an initial randomization through treatment and measurement attrition (Chalk & King 1998, Norman et al 1995, Speer & Newman 1996).

Under these circumstances, it would be comforting to have some assurance that the results of nonequivalent comparison designs were close approximations to those of randomized designs so the inferential strength of the experiment was preserved. Empirical comparisons, unfortunately, give no such assurance. Using meta-analysis summaries, Lipsey & Wilson (1993) found that randomized and nonrandomized designs in the same intervention area often gave quite divergent results. Though the average discrepancy was near zero, they ranged from large positive differences to large negative ones. Similarly, meta-analyses in selected intervention areas have found different results from randomized and nonrandomized comparisons, although the discrepancies were considerably diminished when the studies were equated for such features as amount of treatment to the control group, pretest effect size, selection, and attrition (Heinsman & Shadish 1996, Shadish & Ragsdale 1996). It thus appears that nonrandomized designs often yield biased estimates, but given favorable circumstances or effective statistical control, they are also capable of producing results comparable to randomized designs (Aiken et al 1998, Reynolds & Temple 1995).

Selection Modeling

The frequent bias in nonequivalent comparison designs justifies efforts to improve them. Most of the work on this issue has been directed toward statistical models for selection bias, in particular on ways to statistically control initial differences between intervention and comparison groups to better approximate results from random assignment.

One-stage statistical models, the prototype of which is ANCOVA (Analysis of Covariance), have long been recognized as problematic for adjusting selection bias between groups (Campbell & Erlebacher 1970). However, some analysts have argued that structural equation modeling with latent variables, particularly mean and covariance structure models, are capable of providing good estimates of treatment effects (Bentler 1991). Aiken et al (1994), for example, used this approach to compare the effects of two drug abuse treatments on the daily activities of drug addicts, and Wu & Campbell (1996) applied it to a reanalysis of the Westinghouse Head Start evaluation.

Two-stage models, introduced by Heckman in the mid-1970s, separately estimate the function describing group membership and that describing group differences in treatment outcome, using the predictions of the former as an

instrumental variable in the latter (Heckman 1979). Selection bias may be represented directly in terms of the characteristics differentiating the groups or propensity scores reflecting the probability of group assignment (Rosenbaum 1995), or indirectly with an instrumental variable that is independent of group membership (Newhouse & McClellan 1998). The difficulty of specifying an appropriate instrumental variable and the analytic complexity of these two-stage models has inhibited their use in outcome evaluation despite persuasive advocacy (e.g. Humphreys et al 1996). Nonetheless, recent applications have been made in such diverse intervention areas as resocialization for homeless substance abusers (Devine et al 1997), compensatory education for low-income African American children (Reynolds & Temple 1998), home care under HMO and fee-for-service payment plans (Holtzman et al 1998), and antidepressant drug therapy (Croghan et al 1997).

The major problem with the available statistical approaches to selection bias is the sensitivity of the results to violation of the model assumptions, especially the requirement that all relevant variables be specified (Hartman 1991, Winship & Mare 1992). Much work remains to be done on the question of which models are best for which circumstances (for a good start on this issue, see Stolzenberg & Relles 1997).

Within-Program Variation

Experimental and quasi-experimental designs represent the independent variable as a categorical dichotomy—treatment and control. Within the treatment category there is no differentiation of program components or recognition of any variability in the type or amount of treatment provided to individual recipients. Thus, experimental outcome studies treat the program as a molar whole that is characterized only in terms of whether it is assigned as present or absent for a particular recipient. Looking inside the ''black-box'' of programs, however, reveals various sources of variation, most of which are beyond the evaluator's control, that can have profound effects on the outcomes that are observed. Program services may, for instance, be inconsistently or incompletely delivered, recipients may access or interact with services differently, and the same service regimen may have different effects for different recipients.

Assessments of program effects often do not consider these sources of variation when analyzing outcome data and instead examine only the mean effects for the overall treatment-control contrast. Such analyses produce unbiased statistical estimates of mean program effects if conducted with adequate statistical power, but furnish little information about the variation around those means and its sources. If that variation stems from important program characteristics, however, ignoring it can result in misleading conclusions about what effects the program produced and why. One of the more noteworthy developments in outcome evaluation is recognition that many forms of within-program variation can be crucial to under-

standing program effects. Some of the more significant forms of such variation are described below.

Delayed, Incomplete, or Failed Program Implementation One of the more trenchant lessons from decades of evaluation practice is that intervention programs are characteristically difficult to implement. It may take a new or revised program years to be up and running in the intended manner if, indeed, that is ever accomplished. Moreover, even mature programs may deliver services incompletely, inconsistently, or not at all to some significant portion of their target clientele.

For example, Stecher et al (1994) contrasted residential and nonresidential treatments for dually diagnosed homeless adults. Both interventions involved two phases, the first of which had eight goals (e.g. client engagement, retention, assessment, and treatment planning) and an intended duration of three months. Interviews with administrators and counselors revealed that fewer than 20% of the goals were accomplished for the nonresidential program during its first nine months of operation. By the fifteenth month the implementation reached 75%, where it peaked. Moreover, even though the residential treatment had been operational for four years prior to the study, it was not until the eighteenth month that it attained all eight program goals.

Similar organizational development was discovered by ethnographers studying 14 projects associated with the Comprehensive Child Development Program (CSR Inc. 1997, St. Pierre et al 1997). With the organizational life cycle described in terms of four stages (Shortell & Kaluzny 1988), it was found that projects spent an average of 13 months in the start-up stage, 12 months in the growth stage, 10 months in the stabilization stage, and 6 months as stabilized programs. In addition, there was wide variation around each of these means, and some projects never reached maturity. Even if programs reach a mature stage of implementation, of course, they may not be able to maintain it. Carroll et al (1998), for instance, found modest reductions over time in the intensity of interventions for substance abuse problems. Similarly, McGrew et al (1994) documented ''program drift'' over several generations of the assertive community treatment (ACT) program.

Most programs involve multiple components, activities, or phases such that implementation can be incomplete for all or some elements at any given time. If a program is not implemented, it cannot be expected to produce effects. If distinct phases or components of programs are not implemented, and they are directed toward specific outcomes, it would hardly be a surprise to find no effects for those particular outcomes. If a program is changing over time, either advancing or deteriorating, clients receiving services at one time will be exposed to a different version of the program than their counterparts who experience it at another time. The evaluator cannot safely assume that a program provided any services, the intended services, or the same services to the designated clients and, indeed, can expect to find considerable variability in the nature and amount of services available to the individuals in any group of clients.

Individual Engagement in Services Even if the structural features of a program (e.g. facilities, personnel, management, routine activities) are put in place as planned, and a complete and uniform service package is made available to each target client, it does not follow that a program is well and consistently implemented. Except in cases of mandatory and enforced participation, the target clients must also choose to accept the services and maintain their engagement until the completion of those services (Mowbray et al 1993). It is not at all unusual for program implementation to be severely compromised by low client retention and rendered inconsistent by variability among the target clients in their level of participation.

In one illustrative study that highlighted these individual differences, Maude-Griffin et al (1998) assessed the relative effectiveness of cognitive-behavioral therapy versus a 12-step program for cocaine abusers. Although the treatment plan called for 36 group and 12 individual therapy sessions, participants attended an average of 14 group and five individual sessions; fewer than 15% attended at least three-fourths of both types of sessions. Importantly, the standard deviations were large relative to the means and represented the full range from individuals who attended no sessions to those who attended all sessions.

Not all programs show low rates of service completion or great variability in amount of participation (e.g. Smith et al 1998), but these are nonetheless pervasive problems in many program areas. Moreover, there is not only reason to believe that persons who complete service are systematically different from those who drop out, but that there are differences among noncompleters as well. Kazdin & Mazurick (1994), for instance, found distinctively different client and family profiles for early versus late dropouts from psychotherapy. Evaluators, therefore, will often find that different treatment doses have been received by different types of clients despite a successful effort by the program to make a complete and consistent treatment regimen available to each participant.

Extracurricular Services In open systems like the ones in which social programs are implemented, services are generally available from sources other than the program being evaluated. Some program participants may obtain such services and some of those services may resemble those the program provides or affect, for better or worse, the outcomes the program is attempting to achieve. Occasionally, evaluators have made efforts to assess the receipt of these services (e.g. Carroll et al 1998, Smith et al 1998, St. Pierre et al 1997), usually to determine if they differ between the treatment and control groups. However, little is known about the relationship of various forms of supplementary services to the outcomes of specific types of programs. This is clearly a topic that requires more systematic attention, but also one that involves another potentially significant source of variation among program participants with regard to the nature of the services they experience.

Aptitude by Treatment Interactions Some sources of variation in program implementation can be linked to the propensities of individuals to engage in treatment (or not), seek more or fewer services, or drop out prior to completion of the program. In addition, there is ample evidence in the literature to indicate that individuals with certain characteristics (e.g. high versus low severity of presenting problem) are differentially affected by certain interventions. These aptitude by treatment interactions (ATI) represent a source of variability in program outcomes that can occur even when every individual receives exactly the same treatment (Smith & Sechrest 1991). ATIs are often uncovered through post-hoc analyses to identify subgroups of clients that show larger and smaller intervention effects. Increasingly, however, attempts are made to specify them a priori (e.g. Longabaugh et al 1995, Maude-Griffin et al 1998).

Differentiation of Treatment and Control Conditions The topics discussed above relate to sources of variation within a program and its clientele that may be important for understanding outcomes. In experimental and quasi-experimental studies, however, program effects are observed in the form of differences between the outcomes for the treatment group and those for the control group. If the program is poorly implemented, treatment participants may receive services little different from control participants and have similar outcomes. However, it is also the case that there will be little contrast if the control group receives services comparable to those of the treatment group, even if the latter are well served.

As proposed in Kazdin's (1986) discussion of treatment differentiation in psychotherapy, the service difference between groups represents the relative strength of the intervention. That is, conditions or groups with overlapping service models, leakage of services across purportedly different interventions, and other forms of compensation can reduce the intended contrast between conditions with corresponding reductions in the size of the relative treatment effects.

Given these circumstances, it is somewhat surprising that little attention has been directed at understanding the types and amounts of service available to those in control conditions, especially ''usual care'' controls and other such conditions that are presumed to receive service of some kind. To emphasize this issue, Cordray & Pion (1993) recommended that the nature, amount, and presumed influence of the services available to the control group be characterized as thoroughly as those for the treatment group. In this regard, it should be noted that all the sources of variation within programs and program clientele that are discussed above also potentially apply within control groups as well.

One of the most comprehensive assessments of treatment differentiation was presented by Carroll et al (1998) as part of project MATCH (Matching Alcoholism Treatments to Client Heterogeneity) . Three scales were developed, each depicting the key features of a different manual-based therapy (cognitive behavioral, motivation enhancement, and 12-step facilitation). These were then used to rate the sessions for the three types of therapy according to the degree to which on-type and off-type intervention features were present. Multiple-group profile analysis

showed substantial statistical differentiation between the conditions, although the scale means nonetheless indicated that each therapy was delivered with some features of the other therapies.

Implementation Assessment As indicated above, it is now widely recognized that a program may not actually be implemented as intended, with the result that what is called the treatment condition may be little different from what is called the control condition. Moreover, even when implemented to some meaningful degree on average, there may be enormous variation in the actual treatment delivered to, and received by, different participants. This recognition has resulted in the rise of program description as a component of outcome evaluation, especially description of the extent of program implementation, delivery of services, and treatment dosage received by the intended clientele.

The techniques for examining the nature and adequacy of program implementation serve not only the purposes of outcome evaluators attempting to establish that the intended services were delivered, but also those of formative evaluators investigating ways to improve programs, and managers and sponsors seeking feedback and accountability. As a result, a bewildering variety of overlapping methods and terminology has evolved. The field now recognizes, however indistinctly, the activities of process evaluation (Scheirer 1994), implementation assessment, program monitoring, and management information systems. The most recent addition to this repertoire is performance measurement, a phrase associated with the various government initiatives at the federal and state levels that require agencies to identify their goals and report on their performance in attaining those goals (Hatry 1997, Martin & Kettner 1996).

Implementation assessment as part of outcome evaluation generally requires a multi-level assessment of the organizational functions associated with effective service delivery. These functions can be influenced by external context (e.g. laws, political support, organizational interdependencies), the character of the host organization (e.g. management structures), the nature of the program unit (e.g. required activities, personnel), and the type of services and the clients to whom they are provided (e.g. mode of treatment, nature of client and provider roles). Despite continuing dialogue about the roles of qualitative and quantitative methods in evaluation (Reichardt & Rallis 1994), in practice, multi-method approaches are commonly used for the task of documenting program implementation (e.g. ethnography, surveys, ratings, observations, interviews).

Implementation assessment is typically undertaken to gauge whether a program has reached some, often poorly specified, end state (e.g. ''implemented as planned'') or to support a judgment that the program is sufficiently mature to warrant an outcome evaluation. Fuller explication of what activities and services constitute adequate program implementation is often aided by the creation, refinement, and use of program ''logic models'' (Brekke 1987; Brekke & Test 1992; Cordray & Pion 1993; Julian 1997; Rog & Huebner 1991; Yin 1994, 1997) and ''program templates'' (Scheirer 1996) that depict the program activities in rela-

tionship to each other and the expected outcomes for service recipients. These constructions are usually derived inductively from stakeholders' descriptions of the program-as-intended and information about reputedly successful model programs and best practices.

If systematic descriptions of the intended program have been developed to give guidance, formal measures of program implementation (e.g. treatment strength and fidelity) can be constructed and used as the basis for determining what services were actually delivered and received (e.g. Carroll et al 1998, CSR Inc. 1997, McGrew et al 1994, Orwin et al 1998, Yeaton 1994). The development of such measures requires the same concern about the psychometric properties of the scales as does outcome measurement. Zanis et al (1997) have pointed out that because of low to moderate intercorrelations among service measures and biases in data recording, valid measurement of the type and quantity of services provided to clients can be difficult.

Nonetheless, the systematic assessment of program implementation as a component of outcome evaluation appears to be on the rise. In their review of 359 outcome studies conducted between 1980 and 1988, Moncher & Prinz (1991) found that the use of some form of treatment fidelity assessment increased from 40% to 60% over the course of the decade. Contemporary estimates may be even higher. The majority of these assessments, however, are based on either summary judgments of the overall treatment delivery or studies of adherence to protocol with small samples of clients. As such, the results are mainly indicative of average program implementation levels and do not necessarily speak to the issue of variation in program exposure and participation among individual clients.

A notable characteristic of implementation assessment done in conjunction with outcome studies is that it is most often conducted as a separate and relatively freestanding activity that is not directly integrated into the outcome analysis. Thus, an evaluator will describe the nature and extent to which an intervention is implemented with one set of observations and evidence, then conduct an experimental or quasi-experimental outcome study yielding additional data that is analyzed separately. This is not surprising given that the customary data analysis schemes do not readily permit differentiation of the intervention within the categorically independent variable defined in the experimental design. Nonetheless, fuller integration of information about program process and outcome would likely yield richer insight into the whys and wherefores of program effects. This is a topic to which we will return later in this chapter.

Outcome Variables

The dependent variables in outcome evaluation are measures of the social conditions the program is expected to change. Describing and evaluating such changes, as well as ascribing them to program activity, have generally been problematic in outcome evaluation (Rossi 1997). The problems are fundamental:

knowing what changes to expect, how to measure them, and how to evaluate them.

What Effects to Measure In the world of program practice within which outcome evaluation is conducted, the changes that programs are supposed to bring about are often not well-specified. A major endeavor in evaluation planning, therefore, is the attempt to convert program goals to measurable outcomes. This is an area in which increased use of program theory has made especially useful contributions. By describing the logic that connects program activities to program outcomes, the exercise of specifying program theory identifies the outcomes that can be reasonably expected. Moreover, attention to interim outcomes and mediator variables serves to identify proximal outcomes expected to follow rather directly from program activities as well as the more distal outcomes that the program hopes ultimately to achieve but that may be too remote or diffuse to measure adequately.

Related approaches involve focused investigation of the structure and substance of the outcome domain the program expects to affect. Dumka et al (1998), for instance, described an intensive qualitative inquiry into the nature of parenting stress among three ethnic groups, the results of which were used to develop an appropriate quantitative measure for the effects of an intervention program. In a similar spirit, Cousins & MacDonald (1998) used concept mapping techniques to describe the key dimensions of successful product development projects and thus identify appropriate outcome constructs for management training programs.

Evaluation practice with regard to delineating the intended program outcomes has evolved largely on the presumption that such outcomes are specific to the particular program being evaluated and, therefore, the identification of outcomes must be individually tailored to each program. Owen (1998) has suggested that generic outcome hierarchies, derived from program theory, may contribute the appropriate framework for specifying outcomes for any of a class of programs with similar characteristics that address similar social problems.

How to Measure Outcomes A relative lack of off-the-shelf, validated measures appropriate for many of the social conditions that programs attempt to change continues to cause difficulties for outcome evaluation (e.g. Chalk & King 1998). Moreover, even established measures may be susceptible to distinctive sources of measurement error when used for purposes of assessing treatment effects, e.g. self-serving biases or comprehension problems among low-functioning clients (Lennox & Dennis 1994).

Among the most distinctive measurement issues associated with assessment of intervention outcomes is sensitivity to change. The characteristics that make a measure sensitive to individual differences on a construct of interest are not necessarily the ones that make them sensitive to change on that construct over time (Eddy et al 1998). Although there is general recognition in the field that outcome measures must be sensitive to change, there has been surprisingly little systematic

analysis of the ways sensitive measures can be identified and how sensitivity can be enhanced. Notable exceptions are Stewart & Archbold (1992, 1993), who described seven factors that bear on the sensitivity of a potential outcome measure to change.

Evaluating Effects Given the results of an experimental or quasi-experimental outcome assessment, some determination must be made regarding whether an effect was found and, if so, its magnitude and meaning. Current practice continues to be dominated by statistical significance as the criterion for claiming an effect, despite vigorous and rather convincing arguments that this is a poor and misleading standard (Cohen 1994, Lipsey 1999, Posavac 1998, Schmidt 1996). Various alternatives have been proposed around the theme of reducing the privileged status of the null hypothesis under conditions of low statistical power. These include reliance on confidence intervals rather than point significance testing (Reichardt & Gollob 1997), reporting of the counternull value, that is, the nonnull magnitude of the effect size supported by the same amount of evidence as the null value (Rosenthal & Rubin 1994), and identifying the minimal detectable effect (Bloom 1995).

When an intervention effect is detected by some defensible criterion, the issue of assessing its magnitude and meaning within the context of the program and its goals remains. In current practice this issue is generally neglected; outcome evaluation reports typically say little about effect magnitude other than reporting statistical significance. Various alternatives have been proposed, but none has yet been widely adopted. One approach is to report effect sizes using the statistics developed in meta-analysis for this purpose, e.g. the standardized mean difference effect size or odds ratio (Posavac 1998). Jacobson & Truax (1991) proposed a definition of clinical significance based on attainment of posttreatment outcome scores more similar to those for functional than nonfunctional populations that has received some application (e.g. Ogles et al 1995). Sechrest et al (1996) argued that the measures used to describe psychotherapy outcomes should be calibrated against real behavior and events in people's lives. Many of their suggestions would apply to outcome measures in other intervention areas as well. It is a limitation of the evaluation field that no general consensus has yet emerged about how to handle this fundamental issue.

If we take into account the numerous sources of variability within programs and clients that were discussed earlier, their influence would be expected to produce many different patterns of change across individuals. This variability makes the issue of evaluating effects even more complex. Clearly, a single effect size estimate would not be adequate for summarizing a differentiated set of changes in response to intervention. The statistical methods for assessing individual-level reliable change (either growth or decay) and no-change that have begun to appear in the evaluation literature yield multiple indicators of effects such as the percentage of participants who improved, deteriorated, or exhibited no change (e.g. Speer & Greenbaum 1995).

Elaboration of the Experimental Paradigm

Program outcome evaluation shares a history with agricultural experiments and medical clinical trials in its use of the experimental paradigm to assess the effects of practical intervention regimens. Experimental methods work best, however, when the researcher can (*a*) operationalize the categorical independent variable in a consistent, well-defined manner, such as dispensing known quantities of pills and placebos or types of fertilizers uniformly to all experimental units; (*b*) control the assignment of units to the conditions defined by the independent variable and keep them in those conditions until the outcome data are collected; and (*c*) know what outcomes are important, how to measure them, and how to judge their practical significance.

As the discussion above indicates, it is rare for these circumstances to apply to the evaluation of social programs. Random assignment is often impractical or unsuccessful; treatments are inconsistent; recipients vary in their level of participation, drop out before treatment is completed, and respond differently; and expected outcomes are difficult to define, measure, and appraise. Nonetheless, this situation offers little justification for abandoning the experimental paradigm for outcome evaluation. No viable alternative method has yet been put forward for providing an equally credible answer to the question, ''Does the program work?'' That is, does the program have beneficial effects on the social conditions it seeks to improve? This question is at the heart of outcome evaluation for the good reason that it is what program stakeholders and policymakers generally want to know when they weigh the merits of investing valuable social resources in an intervention program.

What has occurred in outcome evaluation is a steady elaboration of concepts and methods to extend and supplement experimental and quasi-experimental design in ways that better capture and elucidate the often complex relationships between program activities and social change. Many such developments are mentioned above. Two interrelated and especially far-reaching trends in outcome evaluation, however, deserve fuller discussion. These are the increased attention to ''theory'' and the application of new statistical models for the analysis of change.

EXPLANATION AND THE MANY ROLES OF THEORY

Despite the associated practical difficulties and limitations, the experimental method persists as the principal tool for outcome evaluation because it produces the most scientifically credible answers to questions about the effects of social programs on the intended beneficiaries. However, in its basic form it renders essentially a yes or no answer with regard to each outcome investigated. This might well be sufficient if programs were disposable commodities that could easily be organized, implemented, tested, and discarded or replaced if they proved ineffective, and retained otherwise. On the contrary, social programs are politi-

cally, financially, and socially difficult to create and maintain, and it is not realistic to regard them as trials in a trial-and-error problem-solving exercise. Typically, poorly performing programs are expected to improve, at least incrementally, rather than be put out of business entirely. This circumstance puts a premium on learning why an intervention is effective or ineffective so that lessons can be learned and improvements can be made. One thing that has become clear in the evaluation field is that program stakeholders and policymakers want an explanation for the results of an outcome evaluation, not just a statement about whether effects were found on selected outcome variables.

The basic, unadorned experiment or quasi-experiment yields little explanatory information about the effects found or not found. Explaining the results requires a thorough description of the way the program functions and discussion of the relationship between program activity and program effects. In short, the explanation of program effects, or the lack thereof, requires some theory about the way the intervention is presumed to bring about the intended effects, a theory that can serve as a framework for organizing and interpreting information from both descriptive and experimental components of an outcome evaluation.

The relevance of program theory to the evaluation enterprise has long been recognized (e.g. Weiss 1972), but within the last decade it has emerged as a major topic (Rog & Fournier 1997). Despite the high level of interest and discussion, there is no evident consensus on what constitutes program theory or how it should be used (Weiss 1997). Instead, various different images of theory abound, and they are cast in several distinct roles in evaluation practice.

Most common is the use of theory as a planning tool for an evaluation (Julian et al 1995). The program models or logic models derived during evaluability assessment are of this sort (Wholey 1994). Their purpose is to determine whether an agreed-upon conceptualization of the program exists, what it is, and whether it is sensible and feasible. Such logic models typically show the key program activities, the program personnel and clients involved, and the expected results. Once laid out, they often lead to program reconceptualization or refinement, as well as serving to identify questions that might be asked and variables that might be measured in an evaluation.

This form of program theory gives the evaluator a road map that directs attention to what stakeholders view as the critical program activities, the intended outcomes, and the presumed relationships between those activities and the intended outcomes. Its purposes in outcome evaluation are to identify the important variables and relationships that should be studied and to furnish a conceptual framework for organizing and interpreting the results.

A problematic aspect of program theory for these purposes is that it is typically derived entirely from the assumptions and expectations of program stakeholders. As such, its basis is generally clinical experience, informed hunches, and common sense about how certain services might bring about the desired changes in social conditions (Campbell 1986, Chen et al 1997). Its value for organizing the outcome study and interpreting the results, therefore, depends on the extent to which the

views on which it is based represent valid insights about the organizational activities and social change processes that affect the target social conditions.

This circumstance adds another dimension to outcome evaluation, that of assessing the feasibility and plausibility of the program theory itself. Some developments in this direction are beginning to take shape. Rossi et al (1999) presented an extended discussion of the steps and tools needed to evaluate program theories and models. Another effort by Scheirer (1996) centered around the use of program templates as tools for evaluating program content. Although principally descriptive, Scheirer and other contributors to her edited collection urged that the program templates be developed on the basis of ''best'' or ''effective practices.'' Huberman (1996:102) noted that ''careful sifting through research, development, and evaluation literature for the best exemplars to implement locally'' is crucial but labor intensive.

Another perspective views theory as something that should emerge from an evaluation rather than function as a starting point. For example, ethnographic or other qualitative methods can be used to develop ''grounded theory'' (Strauss & Corbin 1990) about how and how well a program works (e.g. Kalafat & Illback 1998). Alternatively, quantitative researchers may engage in exploratory analysis of interactions and relationships to discover informative patterns in the data collected for an evaluation (e.g. Rosenheck et al 1995).

Although either of these forms of program theory might be useful as a framework for explaining program effects, they do not characteristically focus on the web of relationships between program actions and intended outcomes in a conceptually sophisticated way. Often, for instance, these theories stipulate a set of sequenced or chained activities, one of which is presumed to lead to another, without paying any attention to the nature or plausibility of the links that are assumed. Weiss (1997) reported that she could find few evaluations that actually tested any of the causal links implied within the program's chain of activities. Instead, she observed that much of what was done in practice focused on program implementation, e.g. attendance, treatment receipt, and so forth.

The form of program theory that Weiss (1997) found least often is the one that is most pertinent to the challenge of explaining why particular program effects are or are not produced by program actions, namely theory which focuses specifically on the causal mechanisms through which program actions have effects. One variation on this theme is the argument that evaluation research should be theory-driven—that is, it should begin with an articulated theory about how program actions cause the expected effects and organize inquiry around investigation of that theory (Chen 1990, Sidani & Braden 1998). The theories involved may take various forms, e.g. intervening variable theories that examine mediators hypothesized as links between program activity and social outcomes (Donaldson et al 1994), dose-response theories that presume differential outcomes as a function of the nature or amount of treatment contact (Brekke et al 1997), and individual differences theories in which differential outcomes are expected primarily

as a function of characteristics of the recipient (Longabaugh et al 1995, Maude-Griffin et al 1998).

Another variation on the idea of organizing evaluation around a theory of the casual mechanism embodied in the program has been voiced from the perspective of scientific realism (Henry et al 1998, Pawson & Tilley 1997). In this view, programs are assumed to bring about social change through one or more ''generative mechanisms'' operating in the program context. For instance, a program to encourage homeowners to mark their property with identifying numbers might reduce theft through such mechanisms as increasing the difficulty for thieves to dispose of stolen goods or enhancing the detection of offenses by making it easier to establish that someone possesses stolen property (Pawson & Tilley 1997). Outcome evaluation in this scheme is oriented toward determining why a program works (i.e. through what mechanism), for whom, and under what circumstances.

One of the concepts central to discussions of the forms of theory relating to the causal or generative mechanisms in programs is that of change. For these theories, the key issue is explaining how problematic social conditions are transformed by the interaction of the program with those conditions. This concept seems especially promising for further development of the explanatory aspects of outcome evaluation and warrants additional attention.

Describing and Explaining Change

Assessing intervention effects is inherently an investigation of change, in particular, the changes in social conditions brought about by the intervention, which itself represents a deliberate change in the social environment. From this perspective, one of the more interesting trends in outcome evaluation is increased attention to change as an important concept for understanding intervention and its effects. Two domains of change are relevant in this regard, corresponding to the common distinction between program action and the mechanisms through which that action produces social changes (Chen 1990, 1994; Lipsey 1997; Weiss 1997).

First, consideration must be given to the matter of organizational change. Programs represent sets of activities that must be enacted within an organizational context influenced by factors both within and outside the organization. Evaluators investigating whether and why a program has been adequately implemented, or who attempt to assist program managers improve program functioning, must inevitably theorize about organizational change and the factors that inhibit or facilitate it.

The second domain of application for change theories relates to the causal process through which desired changes in social conditions come about as a result of program action. Thus, some set of cause-and-effect links is presumed to connect the intervention to the effects it is intended to achieve. Evaluators investigating whether and why the expected effects occurred must inevitably theorize about social change and the mechanisms that bring it about.

Organizational Change That program implementation represents a form of organizational change was recognized early in the history of program evaluation. Concepts such as formative evaluation (Scriven 1967, 1991), process analysis (Weiss 1972), evaluability assessment (Wholey 1994), and implementation failure (Suchman 1967) highlighted the fact that effective service delivery did not follow automatically from the program status quo. Over the past three decades, recognition of the theories of action underlying programs (Bickman 1987, 1990; Chen 1990, 1994; Weiss 1997), coupled with the development of ways to systematically describe program components and activities (e.g. logic models), has placed organizational issues at center stage in the analysis of program implementation and process.

In addition, the role of the evaluator in relation to program organization has expanded over the decades. In early conceptions of evaluation, evaluators were characterized as ''methodological servants to the experimenting society'' (Campbell 1971), a role largely confined to assessing the effectiveness of programs devised and organized by others. Since then, the evaluator's role has expanded so that it often includes preintervention issues, such as program design and planning, and program implementation and refinement. Thus, evaluation specialists have increasingly infiltrated the organizational development process.

Despite this shift in roles and the heightened interest in program theory, there has been little systematic development of organizational theory focused on the process of implementing social programs. This deficiency is especially striking in light of the availability of a large body of general theory for organizational change that should be adaptable to program implementation issues (Rogers & Hough 1995). Similarly, organizational variables are not generally used to aid the understanding of specific programs undergoing evaluation, though there are notable exceptions (e.g. CSR Inc. 1997, Delany et al 1994).

Aside from involving little organizational change theory, the increasingly systematic efforts in evaluation practice to understand if and how programs are organized for effective delivery of services are often conducted independently of the outcome analyses. Implementation assessment is thus generally a separate, first step in outcome evaluation, but the results of that assessment relating, for instance, to variability in delivery and receipt of services, are not integrated into the analysis of outcome data. Sometimes this is appropriate because the implementation process does not produce or permit any variability, resulting in no individual differences in engagement, retention, and so on (e.g. Smith et al 1998). In many instances, though, better explanation of the observed outcomes would be possible if program implementation data were linked to change on outcome variables for individuals with varying kinds of interaction with the program.

Development and application of forms of organizational change theory for program implementation and service delivery, therefore, hold considerable potential for helping to explain program outcomes. Such theory would also be useful for program planning and formative evaluation aimed at program improvement, even without outcome evaluation. Work along these lines is not very advanced,

however, and currently provides limited tools for program description and analysis.

Social Change The conception of the way in which the intended social improvements come about as a result of program actions that is implicit in a program's activities, or sometimes explicitly articulated, constitutes the program's theory of social change. Such theories depend on assumptions about the etiology of the problems the program attempts to address and the mechanisms by which change can be induced. These theories vary quite dramatically in their level of specificity and applicability across problem domains, though several general theoretical frameworks are routinely encountered as justification for social programming. For example, Smith et al (1998) postulated that problem drinking is the result of environmental contingencies; modifying drinking behavior is a matter of altering the contingencies. Their intervention, the Community Reinforcement Approach (CRA), was derived directly from their theory. Similarly, Latkin et al (1996) used social network theory to craft an intervention to prevent HIV among injection drug users. The guiding principle was that social influence processes have strong effects on risky behavior. Therefore, altering the social processes among members of groups (e.g. social comparison processes, fear of social sanctions, socialization of new members, information exchange) should reduce risky behaviors of all members of the network.

Other meta-theories of change include knowledge-attitude-behavior models (Weiss 1997), in which changes in behavior are presumed to be a function of changes in attitudes, which in turn depend on the acquisition of knowledge (e.g. regarding the harmful consequences of smoking). Similar notions of change underlie interventions stemming from cognitive-behavior theory, e.g. in health (McGraw et al 1996) and substance abuse (Longabaugh et al 1995, Maude-Griffin et al 1998).

The relevance of models of change to social programs is particularly well illustrated by the transtheoretical model championed by Prochaska and his colleagues (Prochaska et al 1992). Although still evolving, the transtheoretical model identifies stages associated with the individual's predisposition to change an adverse behavior (i.e. precontemplation through maintenance of change). This model has become a basis for prevention programs in such areas as smoking (Dijkstra et al 1998), skin cancer (Hedeker & Mermelstein 1998), and substance abuse (see Prochaska et al 1992).

For purposes of this discussion, the transtheoretical change model highlights several important principles about the assessment of change. First, whereas outcome evaluation usually examines the magnitude of the mean change for an intervention group relative to a control group, theories of individual change direct attention to differential patterns of change within groups. By focusing on individual differences in susceptibility to the intervention (stage of change each person is in), we can get closer to the goal of understanding how programs affect individuals, who is most affected, and under what circumstances.

Increasingly, interventions are rooted in theories of change that stipulate individual difference variables presumed to moderate the direction and rate of change. Using these theories, program evaluators have the opportunity to identify and measure these individual difference variables and analyze them in relationship to individual-level change on outcome variables. Combined with information about the nature and amount of participation of each individual in the program, the results of such analysis should yield answers to both the questions of whether the program produced the expected effects and why or why not.

Integrating Theory, Design, and Analysis The fullest and most informative scheme for outcome evaluation would take into consideration (*a*) the variability to be expected in program implementation, service participation, response to treatment, and the like, along with the organizational theory relating to those factors; (*b*) the causal mechanisms presumed to link program action with social change and the moderator and mediator variables associated with that theory; (*c*) the observable outcomes expected to result from the program action at the level of change in the individuals exposed to the program; and (*d*) the net program effects attributable to program action on the basis of an experimental or quasi-experimental design. A careful integration of this information should indicate whether the program brought about change, for whom, why or why not, and in the process, yield useful descriptive information to guide program improvement and general understanding about that particular form of intervention. While individual evaluation studies can be found with one or more of these elements, none combines them all in an integrated fashion. This is due to (*a*) practical constraints—not every evaluation situation affords the resources and opportunity for such a probing inquiry; (*b*) conceptual deficiencies—theoretical development is primitive in many intervention areas, though studies of the sort described would do much to improve it; and (*c*) technical limitations—combining experimental and correlational data in an integrated analysis oriented toward individual and group change issues presents many challenges. On the latter point, however, important and relatively recent statistical developments have equipped evaluators with powerful new tools that can be readily adapted for these purposes.

ADVANCES IN STATISTICAL MODELING

From a statistical point of view, the ''core'' (or initial) analysis of data from an experiment should follow the unit of assignment (Boruch 1997); this is often referred to as the ''intent-to-treat'' model of analysis. Even with many transgressions (e.g. crossovers, dropouts) unbiased estimates of net treatment effects can be derived. As a starting point, Boruch's advice is sound. But, for outcome evaluation also to yield information that helps explain how and why effects were produced, statistical analyses that produce more differentiated results are required.

In addition, theories of change such as those discussed above, when sufficiently well developed, should guide the form and scope of the analysis.

Evaluators have adapted various statistical approaches that have been around for some time to these purposes. The more interesting development, though, is the proliferation of statistical models and approaches to the study of change that has been generated in the last decade by statistical theorists and methodologists (Collins & Horn 1991, Francis et al 1991, Meredith & Tisak 1990, Muthén & Curran 1997, Willett & Sayer 1994). These methods can be grouped into two different analytic strategies. The first focuses on estimating relative group differences. What distinguishes these methods from the intent-to-treat model is the deliberate attempt to incorporate factors associated with program implementation, program theory, or both into the analysis of those group differences. The second analytic strategy is relatively new and focuses specifically on change, generally referred to as ''growth.''

In reviewing practices within each of these analytic strategies, we attempt to distinguish between sequenced and integrated assessments of change. Specifically, some programs involve change processes at several levels of the service system (e.g. community, program, and individual) such that changes at one level are presumed to be prerequisite to change at other levels. These models require a sequential assessment of the steps in the change process (Chen et al 1997). On the other hand, when the change processes all pertain to the same level of analysis (e.g. individuals) and relevant measures are obtained on all change parameters, an integrated model can be used. The most fully integrated model would examine change in the target clientele as a function of variables relating to receipt of service, characteristics of the recipients, and individual-level change processes operating through some sequence of mediating variables. Currently, illustrations of fully integrated models are rare.

Analysis of Relative Group Change

In general, analysis of relative program effects tends to represent change with simple before-after comparisons or, perhaps, longer time series for group means on performance indicators. For example, Smith et al (1998) compared the mean level of drinking behavior over several measurement occasions for program and control clients. Multivariate analysis may be used to link process and before-after change data on outcome variables. Spoth et al (1998), for instance, applied structural equation modeling to examine the direct and indirect effects of family-focused preventive interventions on changes in parenting behavior.

Within many of these studies, examination of program change involves a sequential assessment in which the first step is to determine if the program implementation has reached some criterion level. That is, in the spirit of a ''manipulation check,'' the first stage of the analysis focuses on ascertaining whether the intervention has been applied or delivered as intended at the program or organizational level. If that is established, the analysis then turns to the observed change

among those receiving services. The linkage between program implementation, program theory, and outcomes is chiefly a logical one; if the program was implemented as intended and effects were observed, then the program theory is presumed supported.

In the Smith et al (1998) study, for example, the effects of a Community Reinforcement Approach (CRA) for alcohol-dependent, homeless adults were assessed. A manual-based assessment of adherence to CRA was undertaken, attendance at CRA meetings was recorded, and clients in both the treatment and control groups were questioned about receipt of non-CRA services. This procedure showed that both groups received relevant services, but among CRA clients the overall amount of service was greater, and more of the alcohol-related components of the CRA protocol were received. In the next phase of the analysis, examination of the outcomes revealed change in both groups on the major dependent variables (alcohol consumption, employment, and residential stability), with the relative effects favoring the CRA group only for the alcohol outcomes.

A more integrated analysis involving program implementation variables is achieved by including variables related to individual-level receipt of services and outcomes in the same analytic model. This approach is a variation on the dose-response analysis conducted in studies of pharmaceutical treatments. Program performance is represented in this scheme by indicators of how much service each individual received. The link to outcome is made by examining the relationship between ''dose'' and change on the dependent variables. This link is analyzed correlationally, of course, since dosage is not assigned randomly other than at the level of the global treatment versus control conditions. Babcock & Steiner (1999) presented a useful illustration in the area of domestic violence of what can be learned from this type of analysis, as well as a discussion of the limitations.

In another application, McGraw et al (1996) used multiple regression to show that in a health education program, program fidelity, modifications made during implementation, and teacher characteristics had direct relationships to changes in dietary knowledge, intentions, and self-efficacy among elementary school students. Similarly, Van Ryzin (1996) conducted a path analysis to examine the indirect effects of the type of management of public housing on housing satisfaction among residents. An interesting aspect of this analysis is that it involved a series of mediator and moderator variables derived from an implicit program theory.

As these examples illustrate, it is not uncommon for outcome studies to incorporate variables relating to program implementation or the degree to which recipients are exposed to the intervention and relate them in some manner to the observed change on outcome variables. The program theories implicit or explicit in these analyses, however, are generally no more than assumptions about certain criterion levels a program should reach on key performance indicators before it can be expected to have effects. Little attention has been given to modeling more

complex changes in organization or service delivery over the program life cycle, much less integrating such analysis with outcome data.

Individual Growth Modeling

Although there are important differences among the various approaches, the common theme in growth modeling is a focus on change at the level of the individual unit as the base upon which to construct any other analyses of interest. Using multiwave data on an outcome variable (observed or latent), these models involve at least two levels of analysis. At the first level, the repeated measures within subjects are analyzed, allowing individual change trajectories to be examined directly in terms of their starting values and the rate and shape of change. The second level of analysis compares the individual growth curves to investigate systematic differences among them. A variety of individual growth models, stemming from such fields as biostatistics, education, and psychometrics, have been developed and refined during the last decade (Mellenbergh & van den Brink 1998, Muthén & Curran 1997, Speer & Greenbaum 1995, Willett & Sayer 1994).

It is clearly the more sophisticated multivariate models of change that are most promising for purposes of analyzing change in outcome studies in ways that integrate information about exposure to the intervention, mediator variables, differential responsiveness to intervention among service recipients, and the degree to which effects are maintained over time. These methods are advancing at an impressive rate. In HLM (Hierarchical Linear Modeling) (Bryk & Raudenbush 1992), individual differences in growth or decline are represented by coefficients derived under a random effects model. Osgood & Smith (1995) reported a useful demonstration of how different HLM-based models could test different patterns of treatment effects. For example, treatment effects may be represented in terms of a simple change-from-baseline model, or group effects on growth trajectories with pre-existing factors controlled might be used.

Longabaugh et al (1995) used hierarchical latent growth modeling to assess the interactive effects of individual difference variables, time, and treatment type in an analysis of substance abuse outcomes. Their results are complex but they illustrate the modeling technique, effectively presenting the results using a three-dimensional graphic. Here, treatment fidelity, differentiation, drift, and therapist and site differences were examined prior to the growth modeling (see Carroll et al 1998). Fidelity and treatment differentiation were judged to be sufficient, and these factors were not included in the second phase of the analysis.

Willett & Sayer (1994) translated growth models into covariance structure analysis. Within their framework, predictors and correlates of change can be assessed and extended to between-group effects that represent intervention conditions. Statistical advances for models that prescribe noncontinuous or staged change have also appeared. For example, Hedeker & Mermelstein (1998) developed a threshold-based model of change for *k* stages of change (with a test for

treatment effects and time x treatment effects). They illustrated the model with data from a multi-school skin cancer prevention study.

Muthén & Curran (1997) expanded the random coefficient model beyond a single response variable by incorporating it into a latent variables framework. As a practical guide to growth modeling for intervention studies, their paper is exceptional. In addition to elaborating the random coefficients model, they described a five step analytic process for testing differences between experimental groups and presented methods for estimating statistical power. The recommended steps include developing and testing a separate change model for the control group and the intervention group, testing for differences between those models, analyzing the equivalence of the groups and the effects of initial status on differential growth, and conducting sensitivity analyses on the models. Reanalysis of data from Kellem et al (1994), using these techniques, revealed larger effects, reflecting greater control over extraneous error.

In general, these recent developments (and others reviewed by Muthén & Curran 1997) share much common ground. The main differences revolve around issues of design (e.g. number of measurement waves) and whether manifest or latent variables are used in the analysis. There are also differences among analytic frameworks in their focus on change exclusively or on change plus variables thought to exercise causal influence on the change. What seems clear from the available literature on growth models is that they are sufficiently flexible to be applied to the sorts of program change models that have been described in the evaluation literature.

The evaluation literature, nonetheless, does not yet offer examples of integrated change models that incorporate program, client, mediator, and outcome variables in patterns suggested by program theory. An interesting example of the potential for such integrated analysis, however, is Osgood & Smith's (1995) Boy's Town follow-up study. Among other analyses, they showed how variation in length of treatment affected estimates of change in feelings of isolation, but did not attempt to model the causal mechanisms responsible for that change.

To date, the most complete attempts to incorporate assessments of program implementation into analyses of change among service recipients have followed the two-step logic mentioned earlier, in which implementation is separately checked as a prerequisite to investigating intervention effects. A good example is reported by CSR Inc. (1997), an elaborate implementation analysis of multiple sites associated with the Comprehensive Child Development Program. Using a common model across sites, the researchers assessed implementation at the organizational and individual family levels. Then, with summary indices for key features of the implementation process in hand, St. Pierre et al (1997) conducted growth curve analyses to assess the developmental outcomes for children. Finding no overall effects, they assessed project variation in service levels, along with the prevalence of nontreatment services obtained in both treatment and control groups. This is one of the few studies available that puts together evidence from

the process and outcome evaluation components to assess whether no-effect findings were due to method, implementation, or theory failure.

The advances in statistical modeling have supplied the technical armamentarium to tackle complex evaluation designs that contain controlled and uncontrolled sources of variation. Unlike the intent-to-treat model, which ignores the sources of uncontrolled variation in treatment receipt, nontreatment service seeking, and other individual difference variables, these advanced techniques are more flexible and comprehensive. Yet, despite their advantages, the simple fact is that departures from randomization always leave room for uncertainty. In discussing their results pertaining to treatment exposure, for instance, Osgood & Smith (1995) were clear that this source of variation represents postassignment selection bias. Babcock & Steiner (1999) made the same point in their analysis of the relationship between treatment exposure and subsequent instances of spousal abuse.

As experience accumulates with these more sophisticated methods, we suspect that the next major developments will be in the area of modeling these sources of postassignment selection bias. As described earlier in this chapter, a great deal of attention has been directed toward modeling selection bias in nonrandomized comparisons through the use of propensity scoring and other two-stage selection modeling. With proper measurement of post-assignment selection processes (e.g. propensity to engage in treatment or seek services) additional progress (and controversy) might be made toward exerting some statistical control over these inherently difficult features of experimental methods in application to program outcome evaluation.

CONCLUSION

The strength of experimental methods for outcome evaluation is the scientific credibility with which they answer questions about the effects of intervention on the social conditions they are intended to ameliorate. They answer these questions, however, chiefly in terms of whether there were mean effects on the outcome variables examined and, sometimes, what the magnitudes of those effects were. As valuable and policy-relevant as this information is, it leaves much of the story untold. Knowing what services were delivered and received, what difference that made to the individuals receiving them, whether individuals responded differently, and generally, why certain effects were or were not found is also valuable and relevant.

These additional concerns cannot be easily addressed by experimental methods within the practical and ethical constraints inherent in social programs. Qualitative methods can tell much of this story and, in that regard, are a worthwhile adjunct to even the most comprehensive and rigorous experimental design. Increasingly, however, evaluators are incorporating additional variables into experimental and quasi-experimental designs to acquire more particularistic data about factors that might explain the variability in outcome. For the most part, the scope of these

efforts has been limited and the associated analyses have been adjuncts to the experimental design. The trend, nonetheless, is appropriately in the direction of more fully integrating those variables and issues into the analysis of the experimental comparisons, greatly aided by powerful new statistical techniques.

However informative, statistical modeling of the sources of variability within experimental groups is nonetheless essentially based on correlational relationships. As such, it is only as good as the variables included and the assumptions made about the nature of those variables and their relationships. A notable weakness of the evaluation field in this regard is the paucity of information that has cumulated from decades of outcome evaluation in various intervention domains about the factors involved in selection bias, differential delivery and receipt of treatment, participation in treatment, differential response to treatment, and the predictors of outcome. Without better empirical and conceptual grounding, progress in outcome evaluation methods will continue to lag despite the impressive advances in statistical techniques.

A question critical to the future of program evaluation is whether useful generalization is indeed possible regarding the factors involved in social intervention and the manner in which they interrelate to produce beneficial effects for the target individuals and populations (Adelman & Taylor 1994). The alternative view is that every intervention situation is so distinctive that little useful input can be derived from prior evaluation studies and related research. If each intervention situation is virtually unique, as much evaluation research seems to assume, then there is little hope for either the cumulation of knowledge to aid the development of explanatory models for specific intervention programs or for generalization about what types of interventions are most effective for what types of problems.

There is an important complementarity between the evaluation of individual programs and the strength and completeness of the knowledge base regarding social intervention. The development of explanatory models of program behavior derived through the cumulation and synthesis of empirical findings would allow evaluators to direct their attention to assessing whether programs operate in a fashion consistent with effective practice. This would reduce the necessity of conducting separate, methodologically difficult, outcome evaluations for each individual program whose effectiveness is in question. The effort of designing and implementing high quality social experiments rich in potentially explanatory variables could then be reserved for innovative intervention strategies that are not yet well understood.

LITERATURE CITED

Adelman HS, Taylor L. 1994. *On Understanding Intervention in Psychology and Education.* Westport, CT: Praeger. 279 pp.

Aiken LS, Stein JA, Bentler PM. 1994. Structural equation analyses of clinical subpopulation differences and comparative

treatment outcomes: characterizing the daily lives of drug addicts. *J. Consult. Clin. Psychol.* 62:488–99

Aiken LS, West SG, Schwalm DE, Carroll JL, Hsiung S. 1998. Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: efficacy of a university-level remedial writing program. *Eval. Rev.* 22:207–44

Babcock JC, Steiner R. 1999. The relationship between treatment, incarceration, and recidivism of battering: a program evaluation of Seattle's coordinated community response to domestic violence. *J. Fam. Psychol.* 13:45–59

Bentler PM. 1991. Modeling of intervention effects. In *Drug Abuse Prevention Intervention Research: Methodological Issues,* ed. CG Leukefeld, WJ Bukoski, pp. 159–82. NIDA Research Monograph 107. Rockville, MD: Natl. Inst. Drug Abuse. 263 pp.

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. 116 pp.

Bickman L, ed. 1990. *Advances in Program Theory: New Directions for Program Evaluation,* Vol. 47. San Francisco: Jossey-Bass. 124 pp.

Bloom HS. 1995. Minimum detectable effects: a simple way to report the statistical power of experimental designs. *Eval. Rev.* 19:547–56

Boruch RF. 1997. *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Thousand Oaks, CA: Sage. 265 pp.

Braucht GN, Reichardt CS. 1993. A computerized approach to trickle-process, random assignment. *Eval. Rev.* 17:79–90

Brekke JS. 1987. The model-guided method for monitoring implementation. *Eval. Rev.* 11:281–99

Brekke JS, Long JD, Nesbitt N, Sobel E. 1997. The impact of service characteristics on functional outcomes from community support programs for persons with schizophrenia: a growth curve analysis. *J. Consult. Clin. Psychol.* 65:464–75

Brekke JS, Test MA. 1992. A model for measuring implementation of community support programs: results from three sites. *Comm. Mental Health J.* 28:227–47

Bryk AS, Raudenbush SW. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage. 265 pp.

Campbell DT. 1971. Methods for the experimenting society. Presented at Meet. East. Psychol. Assoc., New York, and Meet. Am. Psychol. Assoc., Washington, DC

Campbell DT. 1986. Relabeling internal and external validity for applied social scientists. In *Advances in Quasi-Experimental Design Analysis: New Directions for Program Evaluation,* 31:67–77. San Francisco: Jossey-Bass. 113 pp.

Campbell DT, Erlebacher AE. 1970. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In *Compensatory Education: A National Debate,* ed. J Hellmuth, 3:185–210. New York: Brunner/Mazel

Campbell DT, Stanley JC. 1966. *Experimental and Quasi-Experimental Designs for Research.* Boston, MA: Houghton Mifflin. 84 pp.

Carroll KM, Connors GJ, Cooney NL, DiClemente CC, Donovan DM, et al. 1998. Internal validity of project MATCH treatments: discriminability and integrity. *J. Consult. Clin. Psychol.* 66:290–303

Chalk R, King P, eds. 1998. *Violence in Families: Assessing Prevention and Treatment Programs.* Washington, DC: Natl. Acad. 392 pp.

Chen H-T. 1990. *Theory-Driven Evaluations.* Thousand Oaks, CA: Sage. 325 pp.

Chen H-T. 1994. Current trends and future directions in program evaluation. *Eval. Practice* 15:229–38

Chen H-T, Wang JCS, Lin L-H. 1997. Evaluating the process and outcome of a garbage reduction program in Taiwan. *Eval. Rev.* 21:27–42

Cohen J. 1994. The earth is round (p<.05). *Am. Psychol.* 49:997–1003

Collins LM. 1996. Is reliability obsolete? A commentary on ''Are simple gain scores obsolete?'' *Appl. Psychol. Meas.* 20:289–92

Collins LM, Horn JL. 1991. *Best Methods for the Analysis of Change.* Washington, DC: Am. Psychol. Assoc.

Cook TD, Campbell DT. 1979. *Quasi-Experimentation: Design and Analysis for Field Settings.* Boston, MA: Houghton Mifflin

Cook TD, Shadish WR. 1994. Social experiments: some developments over the past fifteen years. *Annu. Rev. Psychol.* 45:545–80

Cordray DS, Pion GM. 1993. Psychosocial rehabilitation assessment: a broader perspective. In *Improving Assessment in Rehabilitation and Health,* ed. R Glueckauf, G Bond, L Sechrest, B McDonel, pp. 215–40. Newbury Park, CA: Sage. 334 pp.

Cousins JB, MacDonald CJ. 1998. Conceptualizing the successful product development project as a basis for evaluating management training in technology-based companies: a participatory concept mapping application. *Eval. Prog. Plan.* 2: 333–44

Croghan TW, Lair TJ, Engelhart L, Crown WE, Copley-Merriman C, et al. 1997. Effect of antidepressant therapy on health care utilization and costs in primary care. *Psychiatric Serv.* 48:1420–26

CSR Incorporated. 1997. *Process Evaluation of the Comprehensive Child Development Program.* Washington, DC: CSR Inc. 404 pp.

Delany PJ, Fletcher BW, Lennox RD. 1994. Analyzing shelter organizations and the services they offer: testing a structural model using a sample of shelter programs. *Eval. Prog. Plan.* 17:391–98

Dennis ML. 1990. Assessing the validity of randomized field experiments: an example from drug abuse treatment research. *Eval. Rev.* 14:347–73

Devine JA, Brody CJ, Wright JD. 1997. Evaluating an alcohol and drug treatment program for the homeless: an econometric approach. *Eval. Prog. Plan.* 20:205–15

Dijkstra A, De Vries H, Roijackers J, Van Breukelen G. 1998. Tailored interventions to communicate stage-matched information to smokers in different motivational stages. *J. Consult. Clin. Psychol.* 66:549–57

Donaldson SI, Graham JW, Hansen WB. 1994. Testing the generalizability of intervening mechanism theories: understanding the effects of adolescent drug use prevention interventions. *J. Behav. Med.* 17:195–216

Dumka LE, Gonzales NA, Wood JL, Formoso D. 1998. Using qualitative methods to develop contextually relevant measures and preventive interventions: an illustration. *Am. J. Comm. Psychol.* 26:605–37

Eddy JM, Dishion TJ, Stoolmiller M. 1998. The analysis of intervention change in children and families: methodological and conceptual issues embedded in intervention studies. *J. Abnorm. Child Psychol.* 26:45–61

Francis DJ, Fletcher JM, Stuebing KK, Davidson KC, Thompson NM. 1991. Analysis of change: modeling individual growth. *J. Consult. Clin. Psychol.* 59:27–37

Gueron J. 1997. Learning about welfare reform: lessons from state-based evaluations. See Rog & Fournier, 1997, pp. 79–94

Hartman R. 1991. A Monte Carlo analysis of alternative estimators in models involving selectivity. *J. Bus. Econ. Stat.* 9:41–49

Hatry HP. 1997. Where the rubber meets the road: performance measurement for state and local public agencies. In *Using Performance Measurement to Improve Public and Nonprofit Programs: New Directions for Evaluation,* ed. KE Newcomer: 75:31–44. San Francisco: Jossey-Bass. 102 pp.

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61

Hedeker D, Mermelstein RJ. 1998. A multilevel thresholds of change model for analysis of stages of change data. *Mulivar. Behav. Res.* 33:427–55

Heinsman DT, Shadish WR. 1996. Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychol. Methods* 1:154–69

Henry GT, Julnes G, Mark MM, eds. 1998. Realist Evaluation: An Emerging Theory in Support of Practice. *New Directions for Evaluation* Vol. 78. San Francisco: Jossey-Bass. 109 pp.

Holtzman J, Chen Q, Kane R. 1998. The effect of HMO status on the outcomes of home-care after hospitalization in a Medicare population. *J. Am. Geriatrics Soc.* 46:629–34

Huberman M. 1996. A critical perspective on the use of templates as evaluation tools. See Scheirer 1996, pp. 99–108

Humphreys K, Phibbs CS, Moos RH. 1996. Addressing self-selection effects in evaluations of mutual help groups and professional mental health services: an introduction to two-stage sample selection models. *Eval. Prog. Plan.* 19:301–8

Jacobson NS, Truax P. 1991. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J. Consult. Clin. Psychol.* 59:12–19

Julian DA. 1997. The utilization of the logic model as a system level planning and evaluation device. *Eval. Prog. Plan.* 20:251–57

Julian DA, Jones A, Deyo D. 1995. Open systems evaluation and the logic model: program planning and evaluation tools. *Eval. Prog. Plan.* 18:333–41

Kalafat J, Illback RJ. 1998. A qualitative evaluation of school-based family resource and youth service centers. *Am. J. Comm. Psychol.* 26:573–604

Kazdin AE. 1986. Comparative outcome studies of psychotherapy: methodological issues and strategies. *J. Consult. Clin. Psychol.* 54:95–105

Kazdin AE, Mazurick JL. 1994. Dropping out of child psychotherapy: distinguishing early and late dropouts over the course of treatment. *J. Consult. Clin. Psychol.* 62: 1069–74

Kellem SG, Rebok GW, Ialongo N, Mayer LS. 1994. The course and malleability of aggressive behavior from early grade into middle school: results of a developmental epidemiologically-based preventive trial. *J. Child Psychol. Psychiatry* 35:963–74

Krause MS, Howard KI, Lutz W. 1998. Exploring individual change. *J. Consult. Clin. Psychol.* 66:838–45

Latkin CA, Mandell W, Vlahov D, Oziemkowska M, Celentano DD. 1996. The long-term outcome of a personal network-oriented HIV prevention intervention for injection drug users: the SAFE study. *Am. J. Comm. Psychol.* 24:341–64

Lennox RD, Dennis ML. 1994. Measurement error issues in substance abuse services research: lessons from structural equation modeling and psychometric theory. *Eval. Prog. Plan.* 17:399–407

Lipsey MW. 1997. What can you build with thousands of bricks? Musings on the cumulation of knowledge in program evaluation. See Rog & Fournier, 1997, pp. 7–24

Lipsey MW. 1999. Statistical conclusion validity for intervention research: a significant ($p<.05$) problem. In *Validity and Social Experimentation: Donald Campbell's Legacy*, Vol. I. ed. L Bickman. Thousand Oaks, CA: Sage.

Lipsey MW, Wilson DB. 1993. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *Am. Psychol.* 48:1181–1209

Longabaugh R, Wirtz PW, Beattie MC, Noel N, Stout R. 1995. Matching treatment focus to patient social investment and support: 18-month follow-up results *J. Consult. Clin. Psychol.* 63:296–307

Martin LL, Kettner PM. 1996. *Measuring the Performance of Human Service Programs.* Thousand Oaks, CA: Sage. 138 pp.

Maude-Griffin PM, Hohenstein JM, Humfleet GL, Reilly PM, Tusel DJ, et al. 1998. Superior efficacy of cognitive-behavioral therapy for urban crack cocaine abusers: main

and matching effects. *J. Consult. Clin. Psychol.* 66:832–37

McGraw SA, Sellars DE, Stone EJ, Bebchuk J, Edmundson E, et al. 1996. Using process data to explain outcomes: an illustration from the child and adolescent trial for cardiovascular health (CATCH). *Eval. Rev.* 20:291–312

McGrew JH, Bond GR, Dietzen L, Salyers M. 1994. Measuring the fidelity of implementation of a mental health program model. *J. Consult. Clin. Psychol.* 62:670–78

Mellenbergh GJ, van den Brink WP. 1998. The measurement of individual change. *Psychol. Methods* 3:470–85

Meredith W, Tisak J. 1990. Latent curve analysis. *Psychometrika* 55:107–22

Moncher FJ, Prinz RJ. 1991. Treatment fidelity in outcome studies. *Clin. Psychol. Rev.* 11:247–66

Mowbray CT, Cohen E, Bybee D. 1993. The challenge of outcome evaluation in homeless services: engagement as an intermediate outcome measure. *Eval. Prog. Plan.* 16:337–46

Murray DM, Moskowitz JM, Dent CW. 1996. Design and analysis issues in community-based drug abuse prevention. *Am. Behav. Sci.* 39:853–67

Muthén BO, Curran PJ. 1997. General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychol. Methods* 2:371–402

Newhouse JP, McClellan M. 1998. Econometrics in outcomes research: the use of instrumental variables. *Annu. Rev. Public Health* 19:17–34

Norman J, Vlahov D, Moses LE, eds. 1995. *Preventing HIV Transmission: The Role of Sterile Needles and Bleach.* Washington DC: Natl. Acad. 334 pp.

Ogles BM, Lambert MJ, Sawyer JD. 1995. Clinical significance of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *J. Consult. Clin. Psychol.* 63:321–26

Orwin RG, Sonnefeld LJ, Cordray DS, Pion GM, Perl HI. 1998. Constructing quantitative implementation scales from categorical services data: examples from a multisite evaluation. *Eval. Rev.* 22:245–88

Osgood DW, Smith GL. 1995. Applying hierarchical linear modeling to extended longitudinal evaluations: the Boys Town follow-up study. *Eval. Rev.* 19:3–38

Owen JM. 1998. Towards an outcomes hierarchy for professional university programs. *Eval. Prog. Plan.* 21:315–21

Pawson R, Tilley N. 1997. *Realistic Evaluation.* Thousand Oaks, CA: Sage. 235 pp.

Posavac EJ. 1998. Toward more informative uses of statistics: alternatives for program evaluators. *Eval. Prog. Plan.* 21:243–54

Prochaska JO, DiClemente CC, Norcross JC. 1992. In search of how people change: applications to addictive behaviors. *Am. Psychol.* 47:1102–14

Reichardt CS, Gollob HF. 1997. When confidence intervals should be used instead of statistical tests, and vice versa. In *What If There Were No Significance Tests?*, ed. LL Harlow, SA Mulaik, JH Steiger, pp. 259–84. Hillsdale, NJ: Erlbaum. 446 pp.

Reichardt CS, Rallis SE, eds. 1994. *The Qualitative-Quantitative Debate: New Perspectives: New Direction for Program Evaluation.* Vol. 61. San Francisco: Jossey-Bass. 98 pp.

Reynolds AJ, Temple JA. 1995. Quasi-experimental estimates of the effects of a preschool intervention: psychometric and econometric comparisons. *Eval. Rev.* 19:347–73

Reynolds AJ, Temple JA. 1998. Extended early childhood intervention and school achievement: age thirteen findings from the Chicago Longitudinal Study. *Child Dev.* 69:231–46

Rog DJ, Fournier D, eds. 1997. Progress and Future Directions in *Evaluation: Perspectives on Theory, Practice, and Methods: New Directions for Evaluation* Vol. 76. San Francisco: Jossey-Bass. 111 pp.

Rog DJ, Huebner RB. 1991. Using research and theory in developing innovative programs for homeless families. In *Using Theory to Improve Program and Policy Evaluations,* ed. H-T Chen, PH Rossi, pp. 129–44. New York: Greenwood. 278 pp.

Rogers PJ, Hough G. 1995. Improving the effectiveness of evaluations: making the link to organizational theory. *Eval. Prog. Plan.* 18:321–32

Rosenbaum PR. 1995. *Observational Studies.* New York: Springer-Verlag. 230 pp.

Rosenheck R, Frisman L, Gallup P. 1995. Effectiveness and cost of specific treatment elements in a program for homeless mentally ill veterans. *Psychiatric Serv.* 46: 1131–38

Rosenthal R, Rubin DB. 1994. The counternull value of an effect size: a new statistic. *Psychol. Sci.* 5:329–34

Rossi PH. 1997. Program outcomes: conceptual and measurement issues. In *Outcome and Measurement in the Human Services: Cross-Cutting Issues and Methods,* ed. EJ Mullen, J Magnabosco. Washington, DC: Natl. Assoc. Social Workers

Rossi PH, Freeman HE, Lipsey MW. 1999. *Evaluation: A Systematic Approach.* Thousand Oaks, CA: Sage. 500 pp. 6th ed.

Scheirer MA. 1994. Designing and using process evaluation. In *Handbook of Practical Program Evaluation,* ed. JS Wholey, HP Hatry, KE Newcomer, pp. 40–68. San Francisco: Jossey-Bass. 622 pp.

Scheirer MA, ed. 1996. A template for assessing the organizational base for program implementation. In *A User's Guide to Program Templates: A New Tool for Evaluating Program Content: New Directions for Evaluation.* 72:61–80. San Francisco: Jossey-Bass. 111 pp.

Schmidt FL. 1996. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol. Methods* 1:115–29

Scriven M. 1967. The methodology of evaluation. In *Perspectives of Curriculum Evaluation,* ed. RW Tyler, RM Gagne, M Scriven, pp. 39–83. AERA Monograph Series on Curriculum Evaluation. Chicago: Rand McNally

Scriven M. 1991. Beyond formative and summative evaluation. In *Evaluation and Education: At Quarter Century,* ed. MW McLaughlin, DC Phillips, pp. 18–64. Chicago: Univ. Chicago Press

Sechrest L, Figueredo AJ. 1993. Program evaluation. *Annu. Rev. Psychol.* 44: 645–74.

Sechrest L, McKnight P, McKnight K. 1996. Calibration of measures for psychotherapy outcome studies. *Am. Psychol.* 51:1065–71

Shadish WR, Ragsdale K. 1996. Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *J. Consult. Clin. Psychol.* 64:1290–1305

Shortell S, Kaluzny A. 1988. *Health Care Management: A Text in Organization Theory and Behavior.* New York: Wiley. 524 pp. 2nd ed.

Sidani S, Braden CJ. 1998. *Evaluating Nursing Interventions: A Theory-Driven Approach.* Thousand Oaks, CA: Sage

Smith B, Sechrest L. 1991. Treatment of aptitude x treatment interactions. *J. Consult. Clin. Psychol.* 59:233–44

Smith JE, Meyers RJ, Delaney HD. 1998. The community reinforcement approach with homeless alcohol-dependent individuals. *J. Consult. Clin. Psychol.* 66:541–48

Speer DC, Greenbaum PE. 1995. Five methods for computing significant individual client change and improvement rates: support for an individual growth curve approach. *J. Consult. Clin. Psychol.* 63:1044–48

Speer DC, Newman FL. 1996. Mental health services outcome evaluation. *Clin. Psychol. Sci. Practice* 3:105–29

Spoth R, Redmond C, Shin C. 1998. Direct and indirect latent-variable parenting outcomes of two universal family-focused preventive interventions: extending a public health-oriented research base. *J. Consult. Clin. Psychol.* 66:385–99

St. Pierre RG, Layzer JI, Goodson BD, Berstein LS. 1997. *National Impact Evaluation of the Comprehensive Child Development*

Program: Final Report. Cambridge, MA: Abt

Staines GL, McKendrick K, Perlis T, Sacks S, De Leon G. 1999. Sequential assignment and treatment-as-usual: alternatives to standard experimental designs in field studies of treatment efficacy. *Eval. Rev.* 23:47–76

Stecher BM, Andrews CA, McDonald L, Morton S, McGlynn EA, et al. 1994. Implementation of residential and nonresidential treatment for dually diagnosed homeless. *Eval. Rev.* 18:689–717

Stewart BJ, Archbold PG. 1992. Nursing intervention studies require outcome measures that are sensitive to change: Part one. *Res. Nursing Health* 15:477–81

Stewart BJ, Archbold PG. 1993. Nursing intervention studies require outcome measures that are sensitive to change: Part two. *Res. Nursing Health* 16:77–81

Stolzenberg RM, Relles DA. 1997. Tools for intuition about sample selection bias and its correction. *Am. Sociol. Rev.* 62:494–507

Strauss A, Corbin J. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques.* Thousand Oaks, CA: Sage. 270 pp.

Suchman EA. 1967. *Evaluation Research: Principles and Practice in Public Service and Social Action Programs.* New York: Russell Sage Found.186 pp.

Van Ryzin GG. 1996. The impact of resident management on residents' satisfaction with public housing: a process analysis of quasi-experimental data. *Eval. Rev.* 20:485–506

Weiss CH. 1972. *Evaluation Research: Methods of Assessing Program Effectiveness.* Englewood Cliffs, NJ: Prentice-Hall. 160 pp.

Weiss CH. 1997. Theory-based evaluation: past, present and future. See Rog & Fournier 1997, pp. 41–56

Wholey JS. 1994. Assessing the feasibility and likely usefulness of evaluation. In *Handbook of Practical Program Evaluation,* ed. JS Wholey, HP Hatry, KE Newcomer, pp. 15–39. San Francisco: Jossey-Bass. 662 pp.

Willett JB, Sayer AG. 1994. Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychol. Bull.* 116:363–81

Winship C, Mare RD. 1992. Models for sample selection bias. *Annu. Rev. Sociol.* 18:327–50

Wu P, Campbell DT. 1996. Extending latent variable LISREL analyses of the 1969 Westinghouse Head Start evaluation to Blacks and full year Whites. *Eval. Prog. Plan.* 19:183–91

Yeaton WH. 1994. The development and assessment of valid measures of service delivery to enhance inference in outcome-based research: measuring attendance at self-help group meetings. *J. Consult. Clin. Psychol.* 62:686–94

Yin RK. 1994. Discovering the future of the case study method in evaluation research. *Eval. Pract.* 15:283–90

Yin RK. 1997. Case study evaluations: a decade of progress? See Rog & Fournier 1997, pp. 69–78

Zanis DA, McLellan AT, Belding MA, Moyer G. 1997. A comparison of three methods of measuring the type and quantity of services provided during substance abuse treatment. *Drug Alcohol Depend.* 49:25–32