Ann. Rev. Sociol. 1978. 4:511–32

SOCIAL EXPERIMENTS

◆10564

Henry W. Riecken

School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Robert F. Boruch

Department of Psychology, Northwestern University, Evanston, Illinois 60201

The notion of experimental investigation of social problems is not wholly new in the sociological literature, although it has been substantially developed in recent years. Dodd (1934) is usually given credit as a founder for his study of the effects of a hygiene program conducted in one rural village, while several others were observed as untreated controls. Greenwood's (1945) discussion of the logic of experimental designs is quite formal and abstract, but Chapin's (1947) monograph reports several applications to problems in housing, delinquency, social participation, public welfare, and other areas. Indeed, Chapin's own methodological interest in experimentation goes back even earlier (Chapin 1931, 1938). It is perhaps significant that the studies of Chapin and his students occurred in an era of extensive social reform—Roosevelt's New Deal years—and that active sociological interest in experimentation did not reappear until a quarter century later, again in response to a new wave of societal reforms during the Kennedy and Johnson presidencies.

Chapin distinguished two forms of what would currently be called "quasi-experiments" (namely, "cross-sectional" and "ex-post-facto" designs) from "projected experimental design" by which he meant an approximation to the modern version of social experimentation. The essential features of the contemporary model are: the systematic comparison of the effects of a planned program of social intervention either with no intervention or with one or more alternative active treatments. Generally such comparisons are designed to provide random assignment of participant

0360-0572/78/0815-0511$01.00

units (individuals, clinics, classrooms, etc) to treatments, or at least they incorporate covariance measures for assessing nontreatment effects. Ordinarily pretreatment as well as posttreatment measures of presumed "effect" variables are made. When assignment to treatment cannot be controlled by the experimenter, the design is called a quasi-experiment (Campbell 1969; Cook & Campbell 1975).

Social experimentation is naturally related to questions of public policy toward social problems (Boruch & Riecken 1975). The contemporary version of experimentation grew principally out of the domestic social amelioration emphasis of the 1960s when the federal government launched dozens of major programs aimed at eliminating poverty, raising the condition of minorities, and solving urban problems. Some social program planners realized that information was not sufficient or appropriate to guide the design of certain intervention programs. Accordingly, they opted for controlled trials of miniature versions of plausible alternative programs, rather than staking everything on a single intervention mode that was merely consistent with existing inadequate data.

In contrast to post hoc evaluation studies, then, social experimentation emphasizes the design of the action program itself as the key feature of social intervention, rather than simply the measurement of its outcomes. This relocation of emphasis places a greater burden on social and behavioral theory, for it implicitly demands that attention be paid to the imputed dynamics of the intervention—that is, to the way the program is presumed to work to bring about intended effects.

MAJOR CONTEMPORARY SOCIAL EXPERIMENTS

For practical purposes, one can say that modern social experimentation began in the early 1960s with the Manhattan bail bond experiment and the pretrial conference experiment, both of which were directed primarily by lawyers, not social scientists. Later in the same decade, the massive Negative Income Tax experiment[1] was begun, as well as a variety of smaller scale experiments in treatment of mental illness, education, nutrition, and police patrolling. These and other major experiments up to 1974 were summarized by Riecken et al (1974), and the present review covers the period since that publication.

In that period, additional sites and designs have been added for the experimental study of income maintenance programs, a significant experiment in the effects of police patrolling in a major city has been completed, an experiment on health insurance has begun in three city and three rural

[1]Also known as the Income Maintenance experiment or the Graduated Work Incentive experiment.

sites, and experimental investigations of the effects of housing allowances for poor people are under way in several sites. A brief description of each of these major experiments is appropriate here.

The Kansas City Preventive Patrol Experiment (Kelling et al 1976) was designed to measure the impact of patrolling on the incidence of crime and citizens' fear of crime. The police beats in a portion of the city were allocated to one of three treatments: "reactive," in which no "preventive" patrolling took place, and officers entered the beat area only in response to calls for police service; "proactive," in which the visibility of police patrol was increased to two/three times its usual level by detailing adjacent "reactive" patrollers to join the regularly assigned prowl cars; and "control" beats with the usual one car. The experiment ran for twelve months. Results showed no significant differences among areas in respect to reported crime, victimization reports, arrests, citizens' fear of crime and attitudes toward police, police response time, or several other measures.

The Health Insurance Experiment began enrolling participants in Dayton in 1974 and in Seattle the following year. Two smaller cities and their adjacent rural areas were added in 1976–1977. Sites have been chosen in part to reflect existing variations in the capacity of health care facilities to absorb additional workloads, since one of the objectives of the experiment is to ascertain how much demand for health services will increase with the provision of health insurance at various prices, as well as to learn how providers adapt to increased demand. Participating families are randomly assigned to one of nine insurance plans, which differ in amount of coinsurance from zero (free care) to 95%, since a principal objective is to estimate how varying cost-sharing will affect demand for health services. A third objective is to assess the impact of the various plans upon the health status of individuals and upon the quality of medical care received (Newhouse 1974).

The Housing Allowance Experiments are complementary assessments of the effects upon the housing market of grants to low income families for the purpose of improving the quality of their housing. The impact of such grants upon the demand for improved housing is being assessed in two communities, while the effects upon housing supply are being measured in two other cities.

Somewhat smaller scale experiments are under way to test such social innovations as: supplementation of wages ("supported work") and subsidization of jobs for parolees and released convicts (US Department of Labor 1977), and work release programs (Waldo & Chiricos 1977). Experimental tests of innovative programs and policies have been supported in career education (Gibboney Associates 1977), mental health (Fairweather et al 1969), delinquency (Elliot & Knowles 1976), day care and homemaker services for the elderly (Katz & Papsidero 1977), nutrition and intellectual

stimulation (McKay et al 1977), outcomes of surgical treatments (Chalmers et al 1972), the employment of women on police patrol (Bloch & Anderson 1974), and the effects of television and radio programs on viewers (Minor & Bradburn 1976; Cook & Conner 1976; Searle, Friend & Suppes 1976). A list of over 300 randomized field tests of a large variety of social programs is provided by Boruch, McSweeny & Soderstrom (1977) from economics, political science, psychology, and other behavioral sciences as well as from sociology. Support comes from the private sector as well as federal agencies although the latter are markedly more important in the larger experiments, especially the Departments of Health, Education, and Welfare (HEW); Housing and Urban Development; Labor; and State (AID). Regardless of sponsorship or financial support, most of these smaller-scale experiments have been designed and executed by local staff—usually multidisciplinary —within schools, hospitals, police departments, correctional institutions, and so forth. The major experiments have generally been conducted by research-contracting agencies, both nonprofit and profit, in the nongovernmental sector.

APPROPRIATENESS AND FEASIBILITY

Because social experimentation is novel, often expensive, and generally time consuming, it often encounters objections, especially on the part of program operators and administrators who feel pressure to "solve the problem" instead of "studying it to death," and who ask what justifies experimentation. The most common justification for a randomized experimental test of intervention programs is that it can provide a less biased, less equivocal estimate of program effect than other methods can (Campbell & Boruch 1975; Gilbert, Light & Mosteller 1975). Randomized tests, properly conducted, allow the experimenter to eliminate a variety of competing explanations of the sequelae of a program (Riecken et al 1974). Furthermore, the actualization of an intervention strategy in the form of an operating program always brings to light the implicit premises and unverbalized assumptions that lurk in the conceptual thickets of intervention programs. Operationalizing an intervention means making hundreds of decisions about the way the intended influences are to affect the participants. For example, the original Negative Income Tax experiment had to decide how often to make grant payments to participants—a decision unconstrained by law or custom, but presumably responsive to theory as well as questions of efficiency and convenience (daily payments would have been impractical) and clearly having some bearing upon the way participants were required to manage their funds. Thus an experiment serves to test the feasibility of administering an intervention and helps to perfect its execution. Finally, an experimental trial is a less costly way of discovering ineffective strategies

and misguided interventions than is a national program (Gilbert, Mosteller & Tukey 1976).

The most common objections to social experiments fall into four classes: feasibility, scope, usefulness, and ethicality. For example, critics claim that rigorous experimental designs are virtually impossible to implement in real-world settings and randomization cannot be achieved; or they allege that sophisticated statistical analyses of serial data can substitute adequately for the more expensive experimental procedure. As Boruch (1975b) has pointed out, the first contention is not supported by evidence and the second assumes the existence of appropriate data, a condition that often does not hold. Critics have also leveled charges that experiments unethically deprive members of the control group, make excessive demands on participants, and violate privacy. Boruch replies to these criticisms, as well as to contentions that experiments are unduly insensitive to individual differences among participants, unable to take qualitative information into account, not generalizable, and not useful for program development. As Boruch's discussion makes clear, experimental design and analytic procedures are flexible and diverse enough to meet the technical objections, while many of the alleged constraints on usefulness depend on negotiating and understanding between experimenters and program operators.

The conditions under which an experimental approach to social innovation is appropriate have been discussed by Riecken et al (1974: Ch. II). They mention political, ethical, administrative, and technical considerations, emphasizing the answerability of the question of a program's effectiveness and the likelihood that an answer can actually influence program development, revision, or implementation. Wholey et al (1975) commented on a closely related point, which they characterized as the "evaluability" of a program.

More global issues of appropriateness turn on such questions as: how should programs or policies be selected for experimentation? Who should participate in the design of the experiment, especially in respect to the designation of objectives and the evidence to be sought? How shall the decisions about experimental site and target (participant) population be made? What procedures should be followed in utilizing results and what provisions should be made for independent verification or reanalysis of results? These global issues have been raised in divers quarters but, so far, have not benefitted from systematic attempts to explore them or to arrive at answers.

DESIGN AND MEASUREMENT

Apart from the purely technical issues of specifying treatment, identifying units, minimizing systematic error and achieving randomization, most attention has been given to exploring the several validities of experimental

design. Proponents of experiments have generally stressed their internal validity (accuracy), or the power of experiment to provide unbiased estimates of the true differences between treatments (Campbell & Boruch 1975). Critics have called attention to the limited generalizability or external validity of experiments that are not based on a representative sample of the population of interest (Cronbach 1977). Likewise, the ecological validity of experiments has been questioned, since carefully specified and executed treatments may not be replicated in a routinely operated program that confronts diverse local conditions, employs less well trained operators, gets less careful management, and so forth. Indeed, this last criticism leads logically to the idea of a carefully designed experiment with carelessly implemented treatments as the wisest method for exploring social innovation!

Recent advances in experimentation have investigated the coupling of qualitative with quantitative information, emphasizing multiple ways of knowing and their interlocks (Campbell 1974). Practical methods of combining the two stressed the use of clinical case studies of participants with randomized tests and anthropological or ethnological observation (Fisher & Berliner 1977). Boruch (1975a) has suggested that combining information from experiments with nonrandom data (e.g. time series) collected from other sources may help to close the alleged gap between the internal validity of one approach and the external validity of the other.

Although the means of assuring quality of measurement in social experiments are not special, there seem to be some persistent problems in assessing both response variables and treatment conditions. Readily available standardized measures may well be irrelevant to program objectives or unresponsive to changes the program is attempting to induce (Elinson 1977; Bianchini 1978; Wargo & Green 1978). Nevertheless, the temptation to use well-standardized measures is great because of their established reliability and the probable availability of norms or comparative data.

Some attention has been given to the question of treatment relevance in a different sense. For example, Carver (1975) maintains that because standardized tests are usually constructed to maximize stable individual differences, i.e. high discrimination along a continuum, they are less sensitive to the influence of treatment programs. His illustrations, drawn from reading-test performance, suggest that, at most, about 30% of the variance in performance is accounted for by the treatment. The remaining variance is a reliable and valid indicator of individual differences. Similar results might be found for psychological and sociological measures of an individual's traits. It is apparent that treatment sensitive measures are more appropriate for experiments, and some work has been done in this area. Shoemaker (1975) has proposed a general framework for achievement testing, for exam-

ple, based on the idea that item domain and instructional objectives should be isomorphic. He suggests strategies for developing parallel tests required in repeated measures designs—tests to yield estimates of a response variable that are reasonably precise for the group though not for the individual.

The measures of response chosen in an experiment may be insensitive, owing to ceiling or floor effects, hence poor discriminability that decreases the power of the experiment to detect effects (Minor & Bradburn 1976; Wargo & Green 1978; Elinson 1977). Finally, it has been suggested that criterion-referenced measurement may be the method of choice for maximizing the relevance of a response variable to treatment. There are problems of gauging reliability and validity of tests so constructed, which are discussed by Hambleton & Novick (1973).

Generally, measures of the treatment variable focus on the match between treatment as designed and treatment as delivered in the field, by counting, for example, behavioral acts of teachers in classrooms to determine whether treatment-prescribed regimens are actually being executed (Crawford, Gage & Stallings 1977). A subsidiary question is how well treatments can be implemented (Williams & Elmore 1976). On the other hand, the behavior of the treatment recipient is important—the family that is paid to view a television program may fail to do so while "unencouraged" families may watch at a higher level (Minor & Bradburn 1976); and even ostensibly simple treatments, such as income subsidy payments, may not be fully understood by the recipients even when they actually get the cash (Nicholson & Wright 1977).

Several unusual measurement techniques have been developed to encourage honest answers to questions in "sensitive" areas, e.g. drug use, arrest record, sexual activity, abortion. These usually involve responding randomly to a sensitive or an innocuous question, or otherwise introducing a calculable error into the data, which must then be analyzed by special statistical methods (Boruch & Cecil 1977; Warner 1971; Horvitz, Greenberg & Abernathy 1975).

ANALYSIS

Recent technical developments have advanced the state of the art considerably since 1974. Bock (1975) devised suitable analysis plans for data from repeated measures designs and more recently (Bock & Thrash 1976) developed a better description and a more coherent strategy for the analysis of time-structured data, both cross-sectional and longitudinal. Another area in which advances have been made is in the analysis of imperfectly implemented experiments. Cook & Campbell (1975) developed checklists for identifying threats to internal validity, while Crawford, Gage & Stallings

(1977) worked out procedures for estimating the extent to which control group members may have been accidentally or covertly treated, and for determining how competition between groups may have led to bias in measuring response or implementing treatment. Regrettably, little progress seems to have been made in estimating missing data in experiments, even though the commonly made assumption of random dropout is quite implausible in most experiments, and simple linear models for estimating missing data are often misleading. A promising beginning has been made by Rubin (1977) in devising methods for handling missing data in sample surveys.

Designs intended as randomized experiments are not always reliably executed. Moreover, it will not always be possible to implement a randomized test, and a nonrandomized design, itself imperfectly executed, may be the only available one. Either situation justifies research on analytic techniques that seek to provide an unbiased estimate of treatment effects when units are initially nonequivalent or have been nonrandomly assigned. A fertile field of mathematical development has been enlarging the understanding of how adjustment methods such as covariance analysis, matching, and regression may produce biased estimates of program effect. Articles by Campbell & Boruch (1975), Kenny (1975), Cronbach & Furby (1970), Cronbach et al (1977), and Bryk & Weisberg (1977) shed much light on this disputed topic. Some of the methods have been assayed through the reanalysis of evaluations undertaken by Gilbert, Light & Mosteller (1975), Magidson (1977), and Wortman, Reichardt & St. Pierre (1977). A related topic is the choice of variables used as a basis for adjusting initial differences between treated and control groups, a matter of importance since analyses based on an incomplete set will generally yield biased measures of effect. Indeed, Cronbach (1977) has asserted that specification errors are a far more important analytic problem than unreliability of measures. Deegan (1976) summarized the consequences of including superfluous variables and incomplete variables. Systematic search methods like path analysis and factor analysis can be helpful, but equally plausible models with different variables or different arrangements of variables may fit the data equally well while yielding different estimates of program effects (Magidson 1977). Alternative approaches are illustrated by Sewell et al (1976).

IMPLEMENTATION AND MANAGEMENT

The difficulties encountered in operationalizing a social intervention program and actually executing it in the field have not generated as much interest among scholars as the exotic problems of measurement, design, and analysis; yet correct execution is of the utmost importance for the interpret-

ability of results. It is by now a truism that social experiments are governed by Murphy's Law (Martin 1973), but the literature on management of experiments is scanty. Regrettably, experimenters rarely maintain logs or diaries on problems of management; or, if they do, rarely publish their findings, and much practical knowledge is lost or remains in the oral tradition. Kelling's (1976) paper on managing evaluation staff is a rare and rewarding exception. Pressman & Wildavsky (1973) document the variety of shortcomings and failures encountered in installing a program in a California community. Much can be learned from the case studies of Williams & Elmore (1976) and from their general discussion of analysis and measurement of implementation; but problems of implementation are many and varied, and usually particular to the substance of an experiment, although certain generic problems occur widely.

Contamination of the control group by treatment can occur when program staff are convinced of the value of treatment and morally opposed to "neglecting" the control groups (Mattick & Caplan 1964) or are bored by the inactivity required in the control group (Kelling et al 1976) and find it personally difficult to carry out the roles assigned by the design. Inadvertent contamination can occur, especially when information or persuasive messages constitute treatment, and when adjacency or access to communication channels allows access to controls, thus "treating" them, too. Still a third form of contamination is the attempted self-reclassification by participants who perceive benefits from the experimental treatment and who, by petition, guile, or outright misrepresentation secure for themselves some of the treatment they value.

In most experiments, membership in a control group is an unrewarding experience, usually accompanied by boring requests for information. Despite payment for interviews with controls, the New Jersey Negative Income Tax experiment experienced a much higher dropout in the untreated control group than in even the least remunerative experimental treatment. The Health Insurance Experiment gave up its control group, because it proved too difficult to obtain adequate data on services rendered to controls. Physicians could not be motivated to fill out reports of services for the experiment when their reimbursement did not depend on their doing so.

Perhaps the most pervasive management problems involve keeping treatments consistent with design specifications and achieving accuracy and completeness in data collection. The individuals who administer the treatment will ordinarily adhere to specifications more faithfully if they understand sympathetically the purpose and rationale of the intervention, have adequate incentives to perform, and encounter a minimum of unreasonable obstacles to correct performance. Accordingly, effort spent on motivating the operating staff of an intervention program and making their tasks easy

to execute will be well repaid in adherence to design. To be sure, there may be unexpected practical difficulties in meeting specifications precisely. The managers of experiments must be aware of compromises, willing to make them when necessary, and then to document departures from specifications so that the subsequent analysis can take them into account.

Data collection in an experiment is often repetitious and can become a tedious task, inviting those who actually conduct interviews or administer tests to cut corners. It is hard to overemphasize the need to select, train, and motivate data collectors thoroughly in order to maintain high quality of data, since no amount of statistical manipulation subsequently can repair flaws at the primary data level. Correspondingly, the organization and management of data collected requires a high level of technical skill to obtain files of cleanly edited data, unambiguously identified, that can be easily and cheaply retrieved. Experience suggests that novel and untried systems of data management are always costly, sometimes disastrous.

Much of what is known about implementing experiments in the field is uncodified, is lore or practical experience, and this important but usually slighted aspect of social experimentation is underattended by social scientists.

ETHICAL ISSUES

Granted the basic premise that it is unconditionally unethical to conduct an experiment in which the harm of the treatment outweighs its advantages, the principal ethical issues in experimentation are privacy and confidentiality of information, informed consent, and equity for participants. Only the last is distinctively a problem for experiments in contrast to other forms of social research.

Privacy refers to a state of the individual, i.e. the potential respondent, while confidentiality is a state of the information respondents directly or indirectly provide. Privacy is a matter between questioner and respondent and depends on the extent to which questions and answers are, in and of themselves, intrusive or disturbing. Confidentiality, on the other hand, is a matter of who has access to particular information about a respondent. A respondent may be willing to provide some information under conditions that appear to him to restrict its use appropriately, i.e. to assure confidentiality, but may be unwilling to divulge other information no matter what the conditions of inquiry are, i.e. because the question invades privacy.

Comparatively little is known about what individuals consider private or what the variability in its boundaries is among categories of persons, and little work on the subject has appeared. Instead, effort has been put into techniques of questioning that minimize the invasion of respondent privacy

by concealing answers from the questioner. These techniques are usually based on asking the respondent to apply a randomizing device in order to choose whether he responds to a "sensitive" or an "innocuous" question, e.g. "Have you ever had an abortion?" "Were you born in the month of March?", but replying honestly. (See the section on design and measurement above). Such techniques protect confidentiality as well as privacy, since the questioner does not know which question has been answered.

Considerably more research has been done on methods of preserving confidentiality. The present state of technology for so doing is summarized by Boruch & Cecil (1977). In addition to collecting information anonymously (if repeated measures of the same individual unit are not needed and units cannot be deductively identified by combining characteristics), the experimenter can employ a variety of procedural and statistical tactics such as the use of aliases, microaggregation, "broker" intermediaries, linked files whose key for matching is stored out of reach of subpoena, and innoculation of error at a known rate.

Besides these technical methods, attention has been given to providing testimonial privilege by a general statute to those engaged in social research (Nejelski 1976; American Psychological Association 1976), and such protection has in fact been incorporated in legislation covering research on drug abuse and some vehicular accidents.

Although the federal Privacy Act of 1974 does not address social experiments in particular, some of its provisions have an indirect effect by severely limiting the possibility of linking statistical records from various federal agency archives. Testimony before the Privacy Protection Study Commission (1977) emphasized the chilling effect of privacy law on the conduct of social research and may have led to the Commission's view that statistical research should be treated formally as a special category of archival use, governed by rules that do not hamper scientific research. The US census experiments (Goldfield et al 1977) on the impact of confidentiality statements in interviews suggests that cooperation rates vary markedly with the degree to which confidentiality is assured.

For more than a decade it has been accepted doctrine that experimenters have an ethical responsibility to obtain the informed consent of participants to experimental procedures, the purpose of the inquiry, and the uses to which data will be put. Many investigators have been concerned that full disclosure, while ethically laudible, might have the undesirable consequence of biasing the response to treatment. To be sure, if the experimental treatment were to become established social policy or routine practice, its purpose would surely be knowable by all participants, so this objection seems less forceful in the case of social policy experimentation than it might be for more basic research on spontaneous social behavior. A different concern

about informed consent has begun to develop, however—namely, a growing doubt that consent can ever be fully informed, especially in the case of complex and wholly novel treatments. It is perhaps relevant that discussions of informed consent among members of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research has been increasingly inclined to question the worth of consent procedures and to place more faith in peer evaluation of subject safety in research through Institutional Review Boards. The reviews of the National Commission (1977) on the protection of children have special relevance to this point.

Finally, the random assignment of participants to treatment in an experiment may involve ethical questions of equity in failing to provide treatment to participants in the control group, legal questions of equal protection when federal agencies sponsor experiments, and, indeed, the statutory authority of agencies to do so. The legal issue is not fully settled, although, in two of three cases brought to test HEW authority in this area, judicial decisions upheld the agency's right to assign participants at random (Breger 1976). In practice, the alleged "inequity" of randomization does not seem to upset potential participants when they perceive the procedure as an honest lottery. From an ethical standpoint, randomization may well be the fairest possible way to proceed given a proposed treatment of unknown efficacy.

Contemporary statistical work focuses on designs that reduce or eliminate the risks of depriving a subject of a potential benefit. In the simplest cases, this may involve comparing equally attractive innovations to one another, rather than comparing one or both treatments to a control condition. The more sophisticated approaches combine randomized and nonrandomized tests to minimize the sheer number of individuals who must be randomized, and to capitalize on ambiguity in diagnosis of severity of condition.

Recent analytic work on adaptive sampling in experiments is very promising. The basic objective here is to minimize the number of individuals subjected to risk (or deprived of benefit), while maintaining the ability to estimate the relative effects of treatments (Sobel & Weiss 1970). In a simple variation of the play-the-winner approach, for example, clients are assigned to treatment A until a failure is detected, then subsequent clients are assigned to treatment B. Criteria underlying the use of methods may include: minimizing expected loss of clients assigned to treatment, minimizing the number of clients assigned to poorer treatment with a given total sample size, and unlimited sequential sample (i.e. optimal stopping rules). Investigations now under way focus on relaxing the constraints on these tactics. For instance, they generally assume immediate response to treatment and

a single response variable, and neither assumption is warranted in social experimentation.

Empirical research on the ethics of randomized assignment is fragmented. Willingness of individuals to be randomly assigned reflects a social ethic at least, and Hendricks & Wortman (1975) find in small studies that willingness increases as participation of individuals in the randomization process increases. Other factors, such as the client's perception of rewards and costs of alternative treatment, level of information provided about alternatives, and arguments (moral, scientific, social) for randomization have not been examined.

Gilbert, McPeek & Mosteller's (1977) research on success of innovation is also pertinent. They examined surgical experiments to estimate how often novel treatments were more effective, equally effective, or less effective relative to standards. They found, for some therapies, that the proportion in each category was about equal. This kind of evidence can clarify the equity issue, since it can then be argued that novel treatments are as often harmful as they are helpful.

SOCIETAL CONTEXT OF SOCIAL EXPERIMENTS

Unlike laboratory investigations or the analysis of archival data, social experiments are formulated, executed, and their results used (or not) in an active societal context. That is, from conception to beyond completion, a social experiment is subject directly to the pressures of institutional, political, and constituent interests, which may prevent its initiation, shape its treatments, select the designers, exclude or include sites, influence eligibility rules for participants, accelerate or delay the dissemination of results, magnify or belittle, display or conceal its findings, and use or ignore them in formulating social policy. Many of these issues and others were reviewed in our earlier cited work (Riecken et al 1974; Boruch & Riecken 1975), and were also reviewed by Rossi & Williams (1972), Bennett & Lumsdaine (1975), Bernstein & Freeman (1975), and Weiss (1972). These discussions covered familiar but important themes—for example, the gains and losses accruing from the various organizational and structural relationships that can obtain between experimenters and the sponsors of an experiment as well as between the operators and staff of an experimental program, resistance to experimental findings on the part of those whose interests are threatened by them, the arguments in favor of scientific objectivity on the part of the experimenter vs frank and energetic advocacy of a program he believes in. Freeman & Sherwood (1965) note that program operators and practitioners resist randomization of treatments, because they are reluctant to "deny

services" to the control group and are convinced that they know what kinds of participants will benefit most from the program. Weiss (1975) characterized compactly and clearly the political context in which research results are used when they support the administrators' preconceptions, and when the very decision as to which programs to test is a political decision.

The realization that social experiments necessarily take place within a political and institutional context has become more widespread and better articulated in the last few years. It is now commonplace to recognize that the social policy issues around which experiments are likely to be proposed are ordinarily issues of considerable importance in which the several relevant parties are likely to have substantial interests; and the terms of the experiment no less than its outcomes are likely to involve gain or loss of power, money, prestige, and even group or organizational survival. Withal, the heightened awareness of this aspect of social experimentation has produced little or no increment in well-grounded guidance for increasing the feasibility and maximizing the likelihood of successful experiments.

It has become clearer, however, that the federal government has a large stake in social experiments, not only because of their potential relevance to policy but because the very terms of their design and the questions they attempt to answer can have a direct effect upon agency operations. Furthermore, the interest of the community or social group of participants in an experiment has been clarified. While it had always been recognized that the cooperation of the community at the site was essential for the successful execution of the experiment, less importance had been attributed to the political readiness of the community to act upon its results.

A related matter is the receptivity of the potential participants to innovative programs, which appears to be a significant factor in social experiments even if the treatment is envisioned as having widespread applicability, not limited to the community in which the experimental trial is undertaken. Receptivity may manifest itself in divers ways. Several innovative programs found enrollment of participants more difficult than anticipated. For example, the number of participants in the Housing Allowance Supply Experiment at one site was about half the number expected in the first year of the experiment. The Career Education Experiments, for which participating school districts forecast an oversupply of students, actually had far fewer participants in the first year than could be accommodated. The reasons for underenrollment and slow increase in enrollment are not completely clear. Certain programs such as day care and homemaker services for the chronically ill or services for housebound elderly may grow slowly because the potential participants are not well integrated into the informal communication network of the community, and the usual referral methods (news media and the like) are not adequate to inform and encourage enrollees. In other

instances failure to understand the purpose of the program, to appreciate its benefits may limit enrollment; and sometimes outright skepticism and suspicion about the bona fides of the agency offering the novel program are the source of slow growth.

There does seem to be a "community effect" in some experiments that can accelerate participation in a program (or conceivably halt it) by the dissemination of information about it through informal channels and media considered trustworthy sources by the potential participant group. This effect can, of course, complicate the management of the experiment, especially when two or more variations of a treatment program (including a no-treatment control) are offered to a participant group that shares an information network.

ASSURING QUALITY IN SOCIAL EXPERIMENTS: THE ROLE OF SECONDARY ANALYSIS

The newcomer to the roster of parties interested in experiment is the US General Accounting Office (GAO). Acting under a 1967 Amendment to the Economic Opportunity Act, the Comptroller General began reviewing "the extent to which programs authorized by the Act were achieving the intended objectives" (Marvin 1976). The GAO subsequently expanded its program audit function, established a program analysis division, and has become an important source of post hoc review of all aspects of social experiments supported by federal agencies. The GAO has conducted its reviews in the classical mode of fiscal accounting—that is, by audit of work accomplished rather than by requiring prior review and approval (as the Office of Management and Budget does in "clearing" questionnaires under the Federal Reports Act).

The nature of GAO's interest in auditing experiments is illustrated by the questions its staff raised regarding the Housing Allowance experiments. These had to do with the execution of the experimental design, especially concern about underenrollment in the Supply experiment with the consequent possible impairment of ability to detect market effects of housing allowances, and doubts about the "representativeness" of the sites selected. The GAO auditors apparently considered the possibility that the sites had been chosen because they seemed to offer conditions under which the experimental treatment would have a better than average chance of "succeeding" —i.e. of demonstrating a positive effect. As respondents to this criticism have pointed out, whether it is correct or not—and the evidence is not compelling—the strategy of choosing sites where conditions are optimal for detecting effects of the allowance program is a wholly defensible one. One objective of the experiment is to decide whether a housing allowance is an

efficacious means of improving housing under *any* conditions. If it is not effective under optimal conditions, there is no need for further testing. If it is effective under some conditions, then the question of generalizability becomes relevant.

The clearest conflict of perspectives between auditors and experimenters arose in regard to what they diversely called the "honesty of reporting" or, on the other hand, the "reliability of estimates of" their income by participants in the experiment. The difference in choice of words neatly emphasizes a fundamental disparity in viewpoint. Since eligibility to participate in the allowance program is a function of income, one can understand the source of the auditors' concern about "honesty"; and, since human memory is not infallible, the experimenters' concern for reliability. But the conflict led to a deeper issue when GAO proposed to test "honesty" by having its audit staff conduct reinterviews with participants. Experimenters protested that this procedure would necessarily violate their pledge of confidentiality to participants and would not, at the same time, insure that the second "estimate" would be a more valid one.

The role to be played by GAO audits of federally supported experiments is not yet entirely clear, but the agency undoubtedly has a large responsibility and its participation can have mixed effects. It could, on the one hand, serve to maintain or increase the quality of experimental work and to help assure that data and analysis are not biased deliberately or accidentally. On the other hand, its participation adds a level of review that may well increase the time and the cost of conducting an experiment, even if there is no intent to delay; and it could lead to restrictive regulations or audit practices if there were disagreements on "proper" experimental procedures. Fortunately, this deplorable contingency has been avoided so far. A favorable development has been GAO's request to the Social Science Research Council to establish a committee to examine GAO's role in social experimentation, with a view toward improving its review procedures. The Council has accepted, and a committee is currently at work.

Even before the GAO entered the field, there had been an interest in the reanalysis (secondary analysis) of experiments, stimulated by the interest of the Russell Sage Foundation (Bernstein & Freeman 1975) and others, in the evaluation of social program evaluations. The main purpose of secondary analysis of evaluative data has been establishing the credibility of the original evaluator's conclusions. Other purposes have included testing new analytic methods using the data. Reanalyses for the sake of establishing credibility are exemplified by the Cook et al (1975) analysis of Sesame Street's evaluation; Wortman, Reichardt & St. Pierre's (1977) examination of data from Rand's evaluation of the Educational Voucher Project; Bejar & Rezmovic's (1977) restudy of the Cali, Colombia, experiments in nutri-

tion and childhood education; Magidson's (1977) reanalysis of the Head-start evaluations; and others. Rossi & Lyall's (1976) case study of the Negative Income Tax Experiment dedicates attention to both process of research and quality of earlier conclusions.

Each of the cases cited involves reanalysis of raw records obtained in the original evaluation. This is a rather intensive style of analysis, and alternative approaches have been suggested; e.g. reanalysis based only on summary statistics in a final report. Intensive analysis at the microdata level has proved difficult for a variety of reasons. Locating the data and securing authority for its release is a troublesome problem even for large studies. Poor documentation and inappropriate aggregation are chronic problems. Delays and refusals to accede to requests for data may be due to poor file retrieval practices, or to an unwillingness of the original investigator to disclose data. Recommendations for resolving these difficulties have been presented by Hedrick, Boruch & Ross (1977). They suggest establishing clear control of data by government for government-sponsored projects, clear schedules for release of data and authority for release, improved quality control over documentation, and increased funds for data archiving (see also Bryant & Wortman 1978). Finally a recent review of "metaevaluation" (Cook & Gruder 1978) covers technical problems of secondary analysis exhaustively and suggests methods for handling such problems within the limits imposed by the current state of the art.

SOURCES OF INFORMATION

Information about social experimentation appears in a wide variety of publications, many of them federal government agency reports, but there are several journals and periodicals concerned with program evaluation that carry reports of current developments, including new applications of experimental methods and new solutions to problems of implementing experimental and quasi-experimental designs. In addition, three new organizations have been established to help advance the state of the art.

The new journals include: *Evaluation Quarterly* (Sage), *Evaluation and Program Planning* (Pergamon), and *Evaluation* (Minneapolis Medical Research Foundation). Pertinent new annuals include *Evaluation Studies Review Annual* and *Policy Studies Review Annual.* Special issues of existing journals have also been dedicated to the topic: Bernstein's (1975) issue of *Sociological Methods and Research* covers validity issues, while the Perloff & Perloff (1977) edited issue of *Professional Psychology* covers a variety of topics, including experiments.

Three new organizations have memberships concerned with evaluation in general, including social experimentation. The Evaluation Network pub-

lishes a newsletter on evaluation research and development. The Evaluation Research Society of America focuses on program evaluations and publishes a newsletter. The Council for Applied Social Research is interdisciplinary, dedicated to improving the quality of applied social research in general, and has as a target constituency individuals in government, academia, and the private institutions conducting applied research.

CONCLUSION

It is clear that social experimentation is a particular type of applied social research, carried out in a context of political decision making about the allocation of resources to programs for intervening in societal processes and attempting to ameliorate social problems. As such, experimentation bears an obvious relation to more traditional applications called "evaluation." A word is in order about this relationship as we see it.

Social experimentation represents the next step beyond post hoc evaluation in the deployment of social science to serve practical social ends. Both evaluation and experimentation are concerned with estimating the effects, intended and unintended, of deliberate social interventions undertaken for purposes of improving the condition of members of a society. The principal differences between them arise from (*a*) the placement of a social experimental design at the beginning of an intervention attempt rather than at some point after it has begun; (*b*) the explicit orientation of the intervention as a controlled, randomized trial or test of what is basically a proposition about influencing change in society, conduct, relationships, resources, and so forth; and (*c*) the consequent participation of the experimenter (as a role) in the design of the intervention and its implementation, as well as in the design of a procedure for assessing its effects. Evaluation is a truncated form of social experimentation, in which the evaluator yields to others the responsibility and power for everything except the procedure for assessing effects.

So stated, the ambition of social experimentation may seem large, even excessive. Let us be sure we are understood. The experimenter participates in the design of the intervention; he does not decide or control it as a sole source of wisdom. We do not enthrone social scientist kings, philosopher kings, or any other form of social-planning monarchy. But the experimenter is present at the creation of the intervention, has a maximal opportunity to clarify and understand its premises, to shape its implementation and design a fair test of the proposition it encompasses. The experimenter, under these conditions, can try to create an investigatory design from which something can be learned about the effectiveness of the intervention, instead of being called in after all of the design mistakes have been made and asked to disentangle the inscrutable.

Is such a viewpoint presumptuous? We merely point out that social interventions are launched, without being explicitly designed, every time a law is passed, or an administrative regulation written. Too often, in our view, laws and regulations rest on an infirm basis of intuition, anecdote, folk beliefs, and factually uninformed assumptions. It is surely not unreasonable to suggest that the best available methods of obtaining factual information about the observed outcomes of social intervention should be employed in the construction of national programs; nor an exaggeration to say that experimental design is most likely to provide the most solid information base.

It would be naive, of course, to think that "facts give answers" to social problems, although a former government official once asserted that social analysts do believe that (Matthews 1976). It would be equally naive to think that the mere availability of factual information would depoliticize decisions about programs. But is it idle to suggest that a body of experimentally established findings will at least provide a firm common ground on which interpretive debates can stand?

Advocates of social experimentation have sometimes been accused of preaching an empty formalism, sometimes, of concentrating too narrowly on single and limited acts of intervention while dealing with multiply determined broad-scale problems. In addition to many other contentions that we have dealt with elsewhere (Boruch 1975), these allegations miss the point of the argument in favor of experiments. To put it most simply, the justification for experimenting with social interventions, for trying innovations on a small scale rather than just accepting the intuitions of the politically most powerful and establishing new national programs, seems no different from the justification for pilot testing, trials, or other form of small-scale exploration to find the flaws in a new scheme. The "formalism" simply brings to bear upon the test or trial the best available social research technology for the purpose of maximizing our confidence in the observations of program effect.

Literature Cited

American Psychological Association. 1976. *Report of the Task Force on Privacy and Confidentiality in Psychological Research.* Washington, DC: Am. Psychol. Assoc. 43 pp.

Bejar, I., Rezmovic, V. 1977. *Final report: The secondary analysis of the Cali intervention project.* Methodology and Evaluation Research Report, NIE-OIBR, Evanston, IL: Psychol. Dept., Northwestern Univ. 68 pp.

Bennett, C. A., Lumsdaine, A. A., eds. 1975. *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs.* New York: Russell Sage. 553 pp.

Bernstein, I. N., Freeman, H. E. 1975. *Academic and Entrepreneurial Research,* New York: Russell Sage. 187 pp.

Bianchini, J. 1978. Achievement tests and differential norms. See Wargo & Green 1978

Bloch, P. B., Anderson, D. 1974. *Policewomen on Patrol: Final Report.* Washington, DC: Police Found. 67 pp.

Bock, R. D. 1975. *Multivariate Statistical*

Methods in Behavioral Research. New York: McGraw-Hill. 623 pp.

Bock, R. D., Thrash, W. 1976. Characterizing a latent trait distribution. *Proc. 1976 Dayton Symp. Appl. Stat.* In press

Boruch, R. F. 1975a. Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociol. Methods Res.* 4(1):31–53

Boruch, R. F. 1975b. On common contentions about randomized field experiments. See Boruch & Riecken 1975, pp. 107–142

Boruch, R. F., Cecil, J. S. 1977. *Assuring Privacy and Confidentiality in Social Research.* Res. Rep., Dept. Psychol. Northwestern Univ.: Evanston, IL. 409 pp.

Boruch, R. F., McSweeny, A. J., Soderstrom, E. J. 1977. *Randomized Field Experiments for Program Development and Evaluation: An Illustrative Bibliography (Revised).* Evanston, IL: Psychol. Dept. Northwestern Univ.

Boruch, R. F., Riecken, H. W., eds. 1975. *Experimental Testing of Public Policy, The Proceedings of the 1974 Social Science Research Council Conference on Social Experiments.* Boulder, CO: Westview. 145 pp.

Breger, M. J. 1976. *Legal aspects of social research.* Presented at Symp. Ethical Issues Soc. Sci. Res. Univ. Minnesota, MN (Apr.)

Bryant, F. B., Wortman, P. M. 1978. Secondary analysis: the case for data archives. *Am. Psychol.* In press

Bryk, A. S., Weisberg, H. I. 1977. Use of nonequivalent control group design when subjects are growing. *Psychol. Bull.* 84:950–62

Campbell, D. T. 1969. Reforms as experiments. *Am. Psychol.* 24:409–29

Campbell, D. T. 1974. *Qualitative knowing in action research.* Presented at Ann. Meet. Soc. Psychol. Study Soc. Issues, New Orleans. (To appear in *J. Soc. Issues.*)

Campbell, D. T., Boruch, R. F. 1975. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. See Bennett & Lumsdaine 1975, pp. 195–285

Carver, R. P. 1975. The Coleman report: using inappropriately designed achievement tests. *Am. Educ. Res. J.* 12(1): 77–86

Chalmers, T. C., Block, J. B., Lee, S. 1972. Cont olled studies in clinical cancer research. *N. Eng. J. Med.* 287:75–78

Chapin, F. S. 1938. Design for social experiments. *Am. Sociol. Rev.* 3(6):786–800

Chapin, F. S. 1947. *Experimental Designs in Sociological Research.* New York: Harper. 206 pp.

Chapin, F. S. 1951. The problem of controls in experimental sociology. *J. Educ. Sociol.* 4(9):541–51

Cook, T. D., Appleton, H., Conner, R. F., Shaffer, A., Tamkin, G., Weber, S. J. 1975. *Sesame Street Revisited.* New York: Russell Sage. 410 pp.

Cook, T. D., Campbell, D. T. 1975. The design and conduct of quasi-experiments and true experiments in field settings. In *Handbook of Industrial and Organizational Psychology,* ed. M. D. Dunnette, pp. 223–326. Chicago: Rand McNally

Cook, T. D., Conner, R. F. 1976. Sesame Street around the world: the educational impact. *J. Commun.* 26(2): 155–64

Cook, T. D., Gruder, C. L. 1978. Meta-evaluation Research. *Evaluation Q.* In press

Crawford, J., Gage, N. L., Stallings, J. A. 1977. *Methods for maximizing the validity of experiments on teaching.* Presented at Ann. Meet. Educ. Res. Assoc., New York (Apr.)

Cronbach, L. J. 1977. Remarks to a new society. *Newsletter,* Evaluation Res. Soc. 1:1–2

Cronbach, L. J., Furby, L. 1970. How we should measure "change"—or should we? *Psychol. Bull.* 74:68–80

Cronbach, L. J., Rogosa, D. R., Floden, R. E., Price, G. G. 1977. Analysis of covariance in nonrandomized experiments: parameters affecting bias. In *Occasional Papers of the Stanford Evaluation Consortium.* Stanford Univ., CA 24 pp.

Deegan, J. 1976. The consequences of model misspecification in regression analysis. *Multivariate Behav. Res.* 11:237–48

Dodd, S. C. 1934. *A Controlled Experiment on Rural Hygiene in Syria.* Beirut: Publ. Fac. Arts Sci., Am. Univ. Beirut (Soc. Sci. Ser. No. 7). 128 pp.

Elinson, J. 1977. Insensitive health statistics and the dilemma of the HSA's. *Am. J. Public Health* 67(5):417–18

Elliot, D. S., Knowles, B. 1976. *An Evaluation of the Oakland Youth Work Experience Programs.* Boulder, CO: Behav. Res. Inst. 271 pp.

Fairweather, G. W., Sande s, D. H., Maynard, H., Cressley, D. L. 1969. *Community Life for the Mentally Ill: An Alter-*

native to Institutional Care. Chicago: Aldine. 357 pp.

Fisher, C. W., Berliner, D. C. 1977. *Quasi-clinical inquiry in research on classroom teaching and learning.* Presented at Ann. Meet. Am. Educ. Res. Assoc., New York (Apr.)

Freeman, H. E., Sherwood, C. C. 1965. Research in large-scale intervention programs, *J. Soc. Issues* 21:11–28

Gibboney Associates Inc. 1977. *The Career Intern Program Final Report.* Washington, DC: Natl. Inst. Educ.

Gilbert, J. P., Light, R. J., Mosteller, F. 1975. Assessing social innovations: An empirical base for policy. See Bennett & Lumsdaine 1975, pp. 39–193

Gilbert, J. P., McPeek, B., Mosteller, F. 1977. Progress in surgery and anesthesia: benefits and risks of innovative therapy. In *Costs, Risks and Benefits of Surgery,* eds. J. P. Bunker, B. A. Barnes, F. Mosteller, pp. 124–69. New York: Oxford Univ. Press

Gilbert, J. P., Mosteller, F., Tukey, J. 1976. Steady social progress requires quantitative evaluation to be searching. In *The Evaluation of Social Programs,* ed. C. Abt, pp. 295–312. Beverly Hills, CA: Sage

Goldfield, E. D., Turner, A. G., Cowan, C. D., Scott, J. C. 1977. *Privacy and Confidentiality as Factors in Survey Response.* Presented at Ann. Meet. Am. Stat. Assoc., Chicago (Aug.)

Greenwood, E. 1945. *Experimental Sociology.* New York: Kings Crown. 163 pp.

Hambleton, R. K., Novick, M. R. 1973. Toward an integration of theory and method for criterion referenced tests. *J. Educ. Meas.* 10(3):159–70

Hedrick, T. E., Boruch, R. F., Ross, J. 1977. Policy and regulation for ensuring the availability of evaluative data for secondary analysis. Evanston, IL: Psychol. Dept., Northwestern Univ. 36 pp.

Hendricks, M., Wortman, C. 1975. Reactions to random assignment in an ameliorative social program. Evanston, IL: Psychol. Dept., Northwestern Univ.

Horvitz, D. G., Greenberg, B. G., Abernathy, J. R. 1975. Recent developments in randomized response designs. In *A Survey of Statistical Design and Linear Models,* ed. J. N. Srivastava, pp. 271–86. Amsterdam: North-Holland

Katz, S., Papsidero, J. A. 1977. An experiment in treating chronic illness in the home. East Lansing, MI: Off. Health Services Res. Educ., Mich. State Univ. 37 pp.

Kelling, G. L. 1976. *Development of Staff for Evaluation.* Prepared for Conf. Emergency Med. Services: Res. Methodol., Atlanta, GA. Washington, DC: Police Found.

Kelling, G. L., Pate, T., Dieckman, D., Brown, C. W. 1976. The Kansas City preventive patrol experiment: a summary report. In *Evaluation Studies,* ed. G. Glass, Vol. 1, pp. 605–57. Beverly Hills, CA: Sage

Kenny, D. A. 1975. A quasi-experimental approach to addressing treatment effects in the nonequivalent control groups design. *Psychol. Bull.* 82:345–62

Magidson, J. 1977. Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation: a general alternative to ANCOVA. *Evaluation Q.* 1(3):399–420

Martin, T. L. 1973. *Malice in Blunderland.* New York: McGraw Hill. 119 pp.

Marvin, K. 1976. Research versus decision requirements and best practice of evaluation. In *The Evaluation of Social Programs,* ed. C. Abt, pp. 289–93. Beverly Hills, CA: Sage

Matthews, D. 1976. Speech before the Assoc. Inst. Res., Los Angeles, CA (May 3)

Mattick, H. W., Caplan, N. S. 1964. *The Chicago Youth Development Project.* Inst. Soc. Res., Univ. Mich., Ann Arbor

McKay, H., Sinisterra, L., McKay, A., Gomez, H., Lloreda, P. 1978. Cognitive growth in Colombian malnourished preschoolers. *Science* 200(4339): 270–78

Minor, M. J., Bradburn, N. M. 1976. *The Effects of Viewing Feeling Good.* Natl. Opinion Res. Cent., Univ. Chicago, IL

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1977. *Research Involving Children.* Washington, DC: US Govt. Print. Off. (DHEW Publ. No. 77-0004, 77-0005). 154 pp.

Nejelski, P., ed. 1976. *Social Research in Conflict With Law and Ethics.* Cambridge, Mass.: Ballinger. 197 pp.

Newhouse, J. 1974. A design for health insurance experiment. *Inquiry.* 11(1):5–27

Nicholson, W., Wright, S. P. 1977. Participants' understanding of the treatment in policy evaluation. *Evaluation Q.* 1:2

Pressman, J. L., Wildavsky, A. B. 1973. *Implementation.* Berkeley: Univ. Calif. Press. 182 pp.

Privacy Protection Study Commission. 1977. *Personal Privacy in an Information Society.* Washington, DC: US Govt. Print. Off.

Riecken, H. W., Boruch, R. F., Campbell, D. T., Caplan, N., Glennan, T. K., Pratt, J.

W., Rees, A., Williams, W. 1974. *Social Experimentation: A Method for Planning and Evaluating Social Intervention.* New York: Academic. 339 pp.

Rossi, P. H., Lyall, K. 1976. *Reforming Public Welfare.* New York: Russell Sage. 197 pp.

Rossi, P. H., Williams, W. 1972. *Evaluating Social Programs: Theory, Practice and Politics.* New York: Seminar. 326 pp.

Rubin, D. B. 1977. Formalizing subjective notions about the effects on nonrespondents in sample surveys. *J. Am. Stat. Assoc.* 72:538–43

Searle, B., Friend, J., Suppes, P. 1976. *The Radio Mathematics Project: Nicaragua 1974–75.* Inst. Math. Stud. Soc. Sci., Stanford Univ., CA. 284 pp.

Sewell, W. H., Hauser, R. M., Featherman, D. L., eds. 1976. *Schooling and Achievement in American Society.* New York: Academic. 535 pp.

Shoemaker, D. M. 1975. Toward a framework for achievement testing. *Rev. Educ. Res.* 45:127–47

Sobel, M., Weiss, G. H. 1970. Play-the-winner sampling for selecting the better of two binomial populations. *Biometrika* 57:357–65

US Department of Labor, Employment and Training Administration. 1977. *Unlocking the Second Gate: The Role of Financial Assistance in Reducing Recidivision Among Ex-Prisoners.* (Res. Dev. Monogr. 45) Washington, DC. 207 pp.

Waldo, G. P., Chiricos, T. G. 1977. Work release and recidivism: an empirical evaluation of a social policy. *Evaluation Q.* 1:87–108

Wargo, M., Green, R. 1978. *Minority Group Testing.* New York: McGraw Hill. In press

Warner, S. L. 1971. Linear randomized response model. *J. Am. Stat. Assoc.* 66:884–88

Weiss, C. H. 1972. *Evaluating Action Programs.* Boston: Allyn & Bacon. 365 pp.

Weiss, C. H. 1975. In *Handbook of Evaluation Research,* ed. M. Guttentag, E. L. Struening, pp. 13–26. Beverly Hills, CA: Sage

Wholey, J. S., Nay, J. N., Scanlon, J. W., Schmidt, R. E. 1975. Evaluation: is it really needed? *Evaluation Mag.* 2(2): 89–93

Williams, W., Elmore, R. F., eds. 1976. *Social Program Implementation.* New York: Academic. 299 pp.

Wortman, P. M., Reichardt, C. S., St. Pierre, R. G. 1977. The first year of the education voucher demonstration: a secondary analysis of student achievement. Evanston, IL: Dept. Psychol., Northwestern Univ.