

Social Experiments

HOWARD E. FREEMAN and
PETER H. ROSSI

University of California, Los Angeles;

University of Massachusetts, Amherst

THE PHILOSOPHICAL AND TECHNICAL BASES OF evaluation research and of social experimentation are quite compatible with long-standing traditions in the health field. Given the strong links of the medical field to biological and physical science research, positivism is the leading philosophical stance, and controlled experiments and multivariate analyses are neither unknown nor threatening to medical personnel. Research on the social aspects of medicine can be seen as a natural extension of scientific methods that have been applied with much success to the purely biomedical aspects of health care.

There is also a strong awareness of and concern for the social aspects of medical care, a theme that has its roots in the earliest beginnings of medicine (Rosen, 1979). Indeed, the term "medical sociology" was coined not by a sociologist but by a physician (McIntyre, 1894), writing at the turn of the twentieth century, to designate a broad field of inquiry into the environmental and social structural factors related to disease and into searches for socially effective organizational arrangements for medical care.

These two trajectories explain why evaluation research and social experimentation have been used for a relatively long time in medicine. Indeed, the question may well be raised whether the recent concern

for evaluation and social experimentation in the health field is anything more than simply giving new labels to well-established activities in the health field. It is significant that one of the first social experiments, Dodd's (1934) water-boiling evaluation in the early 1930s, was in the field of public health. During World War II, experiments were undertaken to determine the efficacy of campaigns to increase compliance with efforts to prevent venereal disease among American soldiers (Madge, 1962). Community mental health campaigns were "experimentally" introduced and assessed during the early 1950s (Cumming and Cumming, 1957). Moreover, many community-based clinical trials and biomedical experiments have long included social dimensions of medical care as part of the interventions being assessed (see, for example, Fox, 1959). Finally, it is significant that among the first extensive treatments of evaluation was one written primarily from a public health perspective (Suchman, 1967).

From one viewpoint, evaluation and social experimentation in the health field are hardly new turns in health research. They can be seen as extensions of strongly rooted medical traditions, with perhaps the only innovation being the terms themselves. Undoubtedly a significant portion of the work classified currently as evaluation and social experimentation in the health field, including at least some of the studies that are reviewed in this article, would have taken place in any event.

From another perspective, however, evaluation research and social experimentation may be seen as recent, quite new phenomena, which have emerged from changed moral, social, and political outlooks of the times, and from the cumulative consequences of scientific and technological discoveries in the social sciences. In this sense, recent concern for social experimentation, program evaluation, impact analysis, and the like can be viewed as a broad social movement (Freeman, 1977) that affects human services generally. The social programs of our society have been increasingly subject to scrutiny by means of the instruments of the social science researcher, operating under the broad rubric of evaluation and social experimentation.

Evaluation as Medical R & D or as Part of Health Policy Formation

The older tradition of evaluation and social experimentation in medicine arose out of a "research and development" (R & D) perspective,

in which the main motivation was to improve health care, with the goal of increasing the efficacy of treatments and programs directed at individuals and populations at risk. In contrast, the new social movement toward evaluation research is policy oriented, concerned with such ultimate issues as equity in the delivery of health services, with the proximate goals of affecting public policy with respect to medical services. The new tradition of "evaluation research," then, stems from the greater and greater role played by the government in the provision of medical services.

The R & D perspective is part of the vast medical research enterprise, dedicated to the detection, diagnosis, prevention, treatment, and management of disease. The evaluation of R & D efforts takes place at various levels within the medical enterprise. At one extreme, individuals and households are the targets of programs that seek to find efficacious ways to control or lessen harmful individual or household practices. Examples of such programs are those that seek to lower the incidence and prevalence of substance abuses—alcoholism, drug addiction or smoking—or to instill better health practices such as weight control or the use of dental floss, or to maximize participation in mass preventive efforts such as vaccinations, chest X-rays, or routine examinations for breast cancer.

At the other extreme, the R & D efforts are directed at larger aggregates that include communities and the nation as a whole. Such programs are directed at improving public sanitation, abating noise, or controlling polluting substances.

Some of these studies use health status indicators as outcomes, such as blood pressure, weight, full or partial medical examinations, and self-reports. Others *presume* that the health and illness conditions are related to the outcomes measured, i.e., that taking birth control pills prevents pregnancies or curtailing toxic fumes reduces lung cancer. There also is considerable variation in the causal theories that underlie these experiments. Some interventions are posited on conventional single- and multiple-factor theories (that excessive alcohol use causes automobile accidents), others on the idea that social determinants trigger biophysiological processes (that social stress causes chemical changes that result in arthritic conditions), and still others use host-agent-environment epidemiological models.

Finally, at the margins, it is often difficult to grasp the health-relatedness of the outcome variables. Some mental health impact

evaluations provide sharp illustrations, where social participation, occupational performance, and frequency of sexual intercourse may be used as proxies for psychological conditions. But scepticism and criticism aside, there is a major body of work, continually growing, of social experimental studies on disease phenomena and their control.

The more recent, second perspective on social experimentation reflects "societal" rather than health care concerns and the outcome measures correspondingly are quite different. Motivating these experiments are, for example, societal concerns for more equal access to health care and for controlling the spiraling costs of medical services. These norms are reflected in the range of "access" studies that use such measures as frequency of services, waiting and travel time, and patient satisfaction as outcome variables (Aday et al., 1980); and in intervention efforts to reduce health care costs by shortening length of hospital stay, lowering the unnecessary use of services and of complex and expensive medical technologies. The goals of increasing access and curtailing costs are not necessarily consistent, although they frequently are shared by the same health planning and policy groups.

Corresponding to the difference in goals for this newer tradition in evaluation and social experimentation is a shift in sponsorship. The "new" evaluations tend to be sponsored by agencies that are concerned with the formation of social policy or by foundations or other groups concerned with influencing social policy. Medical researchers and public health specialists may be concerned with the best way to achieve some health-related goal and initiate R & D efforts to find those ways. But the new entrants into evaluation and social experimentation are concerned with public policy and with serving the needs of Congress and the executive branch.

In these policy-oriented studies, the independent variables or interventions range from medical treatments to bureaucratic arrangements for service delivery to practitioner supply and competency. The best known studies are national in scope, and usually are multisite evaluations. In number, however, they are a small fraction of the total evaluation effort: federal, state, and local governments and private philanthropies support many small-scale intervention-evaluation efforts (Aiken et al., 1980). The interventions studied, including most of the national ones, usually are not radically innovative, but represent efforts to increase efficiency in the delivery of services and modestly control the costs of care (Freeman and Solomon, 1979). For example,

in the several federal-private foundation national evaluations currently undertaken the idea at stake is not neighborhood health care centers but whether or not affiliation with medical schools, large tertiary care centers, and so on makes a difference (Robert Wood Johnson Foundation, 1979). Given the nature of the American political system, it is not at all surprising that policy-oriented interventions in the health care system are not radical nor innovative; decision makers are interested in policy alternatives that will make the present system work a bit better and not in totally different systems, whether or not they would work better.

A final distinction between the two perspectives has already been alluded to by the terms "traditional" and "social movement." Studies that directly or indirectly focus on health status as the outcome concern are typically regarded as "insider" evaluations—in the sense that they are seen as properly the domain of health science researchers—and many are undertaken within academic health science settings either by evaluators with identities as health professionals, or by individuals with social science training and a specialization in health and medical care. Perhaps a sharper way of putting it is that most studies related to health status studies are apolitical, not only because the interventions rarely challenge fundamental norms in either the health care field or the social economic order, but also because they are not directed towards changing existing public policy.

Most of the access and cost control efforts, however, require some modification of the system of medical care delivery, who delivers it, and how the worth of the services is judged. Although not often directed at changing things radically, such policy-oriented interventions do challenge at least to a minimal degree the status quo within the medical care system. Hence the new evaluation threatens to some degree the existing medical establishment.

Of course, one should be aware of the dangers of falling prey to one or the other of two oversimplifications. The first is the assumption that all medical care practitioners jealously prize their autonomy, reject regulatory controls, and oppose access and cost-control efforts on loss of autonomy grounds. The second oversimplification is that the self-interests of providers and medical scientists always result in conservative stands on political and social issues in general and particularly on issues related to health care delivery. The history of health

care indicates that some medical scientists and practitioners formulated and advocated innovative positions regarding equality of service entitlement and some were advocates of collective solutions to assure health care would not be an excessive economic burden on individuals.

However, it is fair to remark that mainly "outsiders," ranging from politicians, social planners, labor leaders, and industrialists to academics and "radical intellectuals," have been the leading influences behind efforts to modify access and cost conditions. Certainly, organized medicine, as embodied in the American Medical Association, has had a notoriously poor record as a force to achieve equity of health care at costs affordable to the community.

In short, traditional evaluation efforts are aimed at refining the technical side of medical care. The policy-related evaluations of the past two decades, in contrast, are concerned to change the medical care system in order to achieve socially defined ends (Shortell and Richardson, 1978). Neither type of effort results in radical changes, although the latter challenges more strongly the existing socioeconomic arrangements of the medical care establishment.

What is Social Experimentation?

Elsewhere, we have discussed in some detail the lack of agreement on the boundaries of the field of evaluation research (Rossi et al., 1979). It is unnecessary to repeat the general discussion here, but a consideration of alternative views of the idea of social experimentation is clearly in order.

In one sense, the idea of experimentation is related to the planning, design and implementation of an innovative intervention program. That is, it is the novel characters of the independent or input variables that are the grounds for designating an activity as an "experiment." In another sense, the idea of an experiment is tied to methodological procedures, implying close control over an intervention and randomization of subjects into experimental and control conditions in order to minimize contaminating influences on outcomes.

A case can certainly be made for the view that typically the evaluation research field includes both an innovative intervention and a degree of methodological rigor—particularly a randomized control

group of one sort or another—as necessary requirements of an experiment. The frequently cited Campbell (1969) paper, “Reforms as Experiments,” would substantiate this position.

Semantic difficulties are encountered when innovative interventions are introduced as “experiments” because they are new, but the introduction is not an experiment in a technical sense; or, conversely, when an old, well-established practice is evaluated by use of an experimental design (an experiment in the technical sense). Furthermore, there are limitations to the use of technical experiments, especially in the evaluation of existing practices or well-established programs. Some innovative medical care procedures for a variety of reasons cannot be introduced in technical experiments. Furthermore, as we have indicated earlier, it is often a matter of judgment whether an intervention is different enough from existing practice to be called an experiment in the first (innovative) sense.

Indeed, there are strong limitations on boldness in innovations. The interests of various stakeholders in the medical establishment, the budget constraints facing federal and local governments, and the relative conservatism of American politics all mean that there are limits on what changes can be introduced (Solomon and Freeman, 1979). Although a case can be made that, unless an innovation is quite different from existing practice, it is not likely to make much of a difference in any outcome, one must take into account the fact that the medical care system has undergone considerable improvement over the past few decades and hence additional improvements are increasingly difficult to achieve regardless of the innovation. This seems to be particularly so in current efforts to improve the health delivery system (Aiken et al., 1980; Aday et al., 1980).

In the case of costs, the political realities that surround the pricing of medical care prevail and any proposals that might appear to curtail drastically the incomes of health providers, sharply reduce the autonomy and prosperity of health care institutions, and/or downgrade the stature and power of professional associations and third-party carriers are strongly resisted. To receive support, experiments need to be designed so that they do not tamper too much with the current economic arrangements, or with power and prestige hierarchies.

Thus, there are currently few bold experiments. Perhaps this was always the case. However, in retrospect the 1950s and 1960s appear to have had many more truly exciting innovations or perhaps, as has

been observed elsewhere, programs were just "packaged better" (i.e., promoted as bold and innovative) (Freeman and Solomon, 1979).

At the same time, there is growing interest in the critical examination of what might be thought of as established, well-institutionalized programs. Fiscal conservatism has replaced the liberal financial optimism of the past two decades (as suggested by the passage of California's Proposition 13, and by the proposed Sunset Laws) leading to demands for the assessment of established programs through social experiments.

In this sense, evaluations of established programs can be thought of as social experiments in which demonstration of lack of worth may lead to program termination. Of course, sometimes such evaluation efforts can result in a program being hailed as a marvelous accomplishment. The world-wide effort to eradicate smallpox can only be regarded as a great success since there no longer exists any world-wide threat from this disease! Other evaluation attempts may point to the futility of certain program efforts, such as the repeated failure of mental hospitals to ameliorate the condition of aged, organically ill persons. If reasonably humane board and care facilities can be found in communities, it may be preferable to transfer aged, organically ill persons to such care. Intervening in these sorts of ways not only may save money but also removes vast edifices to our incompetence in dealing with certain human problems.

In evaluating established programs, we often have to abandon adherence to both senses of the term social experiment. After all, a long-established program is hardly an innovation and hence hardly "experimenting" with our institutions. For a variety of reasons, experimentation in a technical sense is often not feasible. Forming control groups means withholding services from some eligible clients, which is often illegal for established programs. Quasi experiments in which self-selected controls substitute for randomized control groups are often possible, but their proper employment requires skillful econometric modeling involving decisions about what are appropriate control populations.

Once quasi experiments are accepted as social experiments, the boundaries between social experiments and epidemiological studies become blurred. In such cases, the requirement that social experiments have an element of experimenter control over the intervention being evaluated—that is, the investigator has to have some control over

implementing some program or preventing some intervention—helps to distinguish between traditional epidemiology and social experimentation. However, “natural experiments,” situations in which changes occur outside the investigator’s control—e.g., changes in the state laws requiring motorcyclists to wear helmets (Muller, 1980)—probably should not be ruled out as defining social experiments. Specific instances involving statistical investigation of the effects of naturally occurring variations in treatment (e.g., Cutright and Jaffe, 1977) can easily be regarded as either epidemiology or as social experiments, depending on whether or not one regards the naturally occurring variation as providing the appropriate setting for a “natural experiment.”

Because the boundaries between social experimentation and other forms of applied social research are so indefinite, we have not drawn strict and necessarily arbitrary limits on the studies we shall consider in this article. Indeed, the specific studies analyzed here have been selected to provide a broad spectrum of work in the field, representing important and interesting experiments in terms of the dual perspectives as well as a range of subject matter and methodology. As a set, they provide convincing evidence of the broad scope and high activity level of evaluation research and social experimentation in the health field.

“True” Experiments: Randomization and Tight Controls

For the applied social research aficionados, the most interesting problem is that of estimating effectiveness of treatments, and for that purpose there is no more desirable research design than the classical (“true”) randomized experiment. Although most commentators agree that such designs provide the best opportunities for effectiveness estimates, they disagree widely over how frequently it is possible to undertake randomized experiments.

The reasons pro and con on the feasibility of controlled experiments will not concern us here in any great detail. Some arguments center around equity and ethical issues that have to do with the randomization of individuals into experimental and control groups. Others have a more practical base, revolving around the acknowledged difficulties

of maintaining experimental and control groups intact over relatively long periods of time, or concerning the high costs of randomized experiments, or concerning the effects of self-selection in experiments where subjects have to volunteer to participate in the experimental treatments, and so on. The fact of the matter is that there appear to be circumstances in which true experiments are not only feasible but highly appropriate, and other circumstances in which it is difficult, impractical, or perhaps unethical (at least in some lights) to undertake experiments (Bennett and Lumsdaine, 1975; Riecken and Boruch, 1974).

Certainly true experiments have been undertaken in a wide variety of social fields. A recent review (Boruch et al., 1978) provides ample documentation of this point, referencing several hundred randomized experiments covering virtually all human service fields, including health services. Experiments in the health field have ranged in scale from the very modest to quite elaborate, as the illustrations given below indicate.

Modest True Experiments

Small-scale randomized experiments are particularly suited for the development and testing of new methods in health care. Indeed, they function primarily to demonstrate the possibilities of a new method. Two prime examples are given below, each illustrating the feasibility of experimentation but neither demonstrating that the methods involved could be used on a large scale. In short, as Campbell and Stanley (1966) assert in their pioneering exposition of the virtues of randomized controlled experiments, such researches are especially powerful because of their "internal validity," their ability to establish the validity of causal relationships, but are limited in generalizability.

Skipper and Leonard's (1968) study of maternal stress and successful outcomes of children's hospitalization is an example of comparatively simple controlled experiments in health care, testing the proposition that the anxiety levels of individuals and significant others is important to health care outcomes. Sets of child tonsillectomy patients (three to eight years of age) and their mothers coming to a hospital were randomly divided into experimental and control groups. The mothers of the 40 patients in the experimental group were exposed to a program of information and counseling with a hospital staff member; the

mothers of the 40 patients in the control group were handled as was usual for the hospital in question. Differences between experimentals and controls were in hypothesized directions. For example, postoperative systolic blood pressure was 17 points higher for the controls than for the experimentals; mean postoperative pulse rates for the two groups were 122.2 and 101.6, and mean postoperative temperatures were 100.7 and 100.1. Other measures of outcome such as vomiting and mothers' responses to a posthospital follow-up questionnaire confirmed the positive effect of the intervention. A rather productive little experiment with clear implications for how we train providers and socialize patients and their kin in anticipation of surgery; Skipper and Leonard (1968:286) also modestly suggested implications for social psychological theory.

While Skipper and Leonard have demonstrated by their relatively inexpensive experiment that, through the use of a seemingly simple intervention directed at the mothers, it is possible to materially improve the postoperative conditions of young children who undergo minor surgery, the experiment does not lead to the conclusion that it would be possible to introduce this method with the same results in a variety of hospital settings. What works quite well in an experiment supervised by two dedicated researchers, in a hospital whose administration and staff may be especially receptive to the theories that underlie the intervention in question, might be ineffective in the more usual American hospital whose staff may be less receptive to such ideas and perhaps overburdened with the tasks of regular hospital duties. A specially engineered trial of a medical care method may be difficult to turn into a production run, i.e., to be incorporated at the same strength and content within regular hospital procedures.

A second example of a modest scaled experiment (Lewis and Resnick, 1967) is much more of a "policy" study. When undertaken, it was somewhat radical or at least not exactly consistent with the ethos of organized medicine. Currently, but even more so in the 1960s, aging patients with chronic illnesses requiring recurring ambulatory care represent a continual strong demand on the resources of hospitals. Realizing that the costs of physician time was an important element in the resource drain by such patients, the experimenters sought to test how well the less costly medical personnel (nurse practitioners) could serve such patients, and at what alteration in the quality of medical care furnished. Indeed, there was some reason to expect that

medical care quality might not suffer at all but even improve since physicians have often been observed to express some considerable impatience with geriatric patients.

The experimental design was a simple one: 66 older patients were first stratified on diagnoses of chronic conditions, sex, age, and race, and then randomly assigned to either physicians or nurse-practitioners. Because of the stratification, no major health differences existed between the experimentals and controls. However, in a pre-experimental interview all patients showed decided preferences for doctors rather than nurses as providers of medical care. At the end of a year of the experiment, the 33 nurse-clinic patients had made 345 visits to the clinic, 95 percent without physician consultation; the 33 in the control group had made only 153 visits. Only 2 nurse-practitioner patients had dropped out, and doctor preference scores declined by about 80 percent. Perhaps most important, the nurses' patients had only 25 percent of the hospital encounters of the doctors' patients. Moreover, average total costs of care per year were estimated at \$98.51 for the nurse-clinic group and \$127.24 for the doctor-clinic one. Although additional studies have raised questions of the comparative patient costs of "new" health practitioners (mostly because of increased and longer encounters) (Spector et al., 1975), the study discussed here, and others of a decade or so ago, clearly have led to modified policies concerning the requirements of primary-care provider resources.

Note that, as in the case of the Skipper and Leonard experiment discussed earlier, this experiment does not "prove" that nurse practitioners do as well as doctors in providing medical care. It does show that under some circumstances they can. Whether nurse practitioners in other hospital settings can also achieve as good results on the average is problematic. However, as the diffusion of this health care method has indicated, this experimental demonstration was strong enough evidence to invite imitation in practice.

Given the current concern with the usefulness and applicability of applied social research, the sequelae to the Lewis and Resnick (1967) experiment are interesting. The specific influence of particular studies is difficult to assess, but it is clear not only that a number of "new practitioner" models have developed and a variety of training programs have emerged, but also that a remarkable spate of research papers has appeared, evaluating and commenting on these efforts. An incomplete literature survey reveals at least fifteen in the past six years

(Bessman, 1974; Brown et al., 1979; Burkett et al., 1978; Chambers et al., 1978; Connelly and Connelly, 1979; Ford and Silver, 1967; Garfield et al., 1976; Levine et al., 1978; Pesznecker and Draye, 1978; Simborg et al., 1978; Spector et al., 1975; Spitzer et al., 1974; Spitzer et al., 1976a; Spitzer et al., 1976b; Sullivan et al., 1978).

Large-Scale True Experiments

Large-scale true experiments, that is, multisite interventions with "national" implications, of course, are rare, in part because of the costs and logistic difficulties of conducting them and, in part, because the rationale for most such efforts is their policy implications and those who promote them are typically unwilling to wait the extended period of time required for the completion of prospective evaluations. (Social experiments on national issues on a large scale have been conducted on income maintenance policies, housing allowances, supported work, and the extension of unemployment insurance benefits to ex-felons released from state prisons.) But in the health field, as in others, when the pressure for knowledge is great enough, in the face of high economic and social costs of inappropriate actions, support emerges for experiments.

Rand's study of alternative national health insurance programs, currently under way, is an illustration, and is perhaps the most exciting and most costly experiment in medical care. Because the study is expected to provide information that would aid in federal policy-making, it could not be conducted in only one or a small number of sites. Rather, the experiment had to be either on a national scale or within a sufficient number of sites of varying location and characteristics to appear reasonable as representing the conditions to be found in a national program. In addition, the experiment had to test out a variety of national health insurance plans, which must cover those policy alternatives likely to be considered by Congress and the executive branch.

This study has a number of important policy objectives:

1. To estimate how alternative cost-sharing arrangements affect demand for health care services.

2. To assess the effect of varying the costs of health services on the health status of individuals.
3. To determine whether cost-sharing arrangements have a different effect on low-income families than on higher-income families.
4. To ascertain how the ambulatory care system accommodates to varying levels of health care demand.
5. To learn about difficulties in administering health insurance plans that place ceilings on out-of-pocket payments by a family.
6. To study the relations between quality of care received and the insurance plan covering the family.
7. To compare utilization, quality of care, health status outcomes, and consumer satisfaction in prepaid group practice and fee-for-service patient groups.

In order to accomplish the objectives of this ambitious evaluation, a total sample of some 8,000 individuals in over 2,700 families were enrolled from four communities—Dayton, Ohio; Seattle, Washington; Pittsfield, Massachusetts; and Charleston, South Carolina. Families were enrolled into the study for either three or five years, and assigned to one of five different plans or to control conditions under which “normal” health care financial arrangements prevailed.

Given the continual legislative press for some form of national health insurance, and the clear commitment of the leadership of both political parties to some such scheme, it is evident that major reasons for delays in implementing national health insurance are disputes regarding the kinds of coverage to be offered and the consequences of resulting different incentives and disincentives to seek health care services.

The still continuing national health insurance study clearly can and will provide valuable answers to key policy questions. At the same time, it is worth emphasizing that such studies are undertaken infrequently and under the pressure for knowledge related to major and highly controversial policy issues. To accomplish such a study requires not only the usual research competencies, but also individuals who are masters at the many administrative and logistical activities required. They must also be skillful both politically and in terms of community relations, for such evaluations require the long-term support of a variety of different influence groups, as well as of the participating families themselves.

Further, such an effort must begin with rapid yet rigorous development of instrument and measurement procedures. The tasks involved in devising appropriate measures and data collection approaches are not to be underestimated. If large-scale studies suffer in execution and accomplishment, the reason is insufficient time and resources to undertake fully the necessary methodological planning and instrumentation. The national health insurance study is a case where there was some lead time, although most would say not enough. In many cases, the demands for information to be taken into account in policy decision-making require so rapid an implementation of evaluations that sacrifices in the state of the art must occur. The need for anticipatory experimentation and an early intelligence system has been suggested numerous times (Bernstein and Freeman, 1975), and experiences such as the national health insurance strain illustrate why such a recommendation makes sense.

At least up to this year, the implementation of large-scale social experiments and evaluations has not been particularly hindered by lack of funding. At a federal level, expenditures for such programs as Medicaid obviously are huge, and using 1 or 2 percent of the total funds for evaluation provides a generous pool to support studies such as the massive experiment by Newhouse and his associates at the Rand Corporation.

It should be pointed out that the issues that the Rand study addresses have been at the forefront of work in the health services research field for decades. The uniqueness of the Rand study lies both in its national scope and in the fact that it is a true experiment with a controlled intervention. The recent four-volume compendium of health services research issued by the National Center for Health Services Research (Freeburg et al., 1979) includes a large number of efforts to examine utilization, satisfaction, compliance, and costs in relation to variations in insurance provisions. Although these studies have provided considerable useful information, with relevance to policy formulation, they tend to be limited both by the characteristics of the study groups examined, and by the various constraints associated with undertaking retrospective quasi-experimental analyses, as we shall describe below.

Large-scale true experiments are jewels, especially to researchers to whom their technical qualities have special appeal. As instruments

of policy formation, however, their utility has been limited. First, as emphasized again and again in this paper, they are quite expensive, even though probably less costly than the large-scale errors in policy they are designed to avert. For example, the total costs of the innovative housing allowance experiments have now reached \$134 million and show some promise of consuming perhaps an additional \$10 to \$20 million before being totally finished. Second, the experiments have so far been limited largely to income transfers—health insurance, housing allowance payments, unemployment benefits, and the like—all treatments that are relatively easy to deliver in a fairly uniform way. Whether or not large-scale experiments can be carried out as well with treatments that comprise labor-intensive human services is problematic. Third, the experiments have all been plagued with design maintenance problems. Participants drop away over time or refuse to participate. For example, only 60 percent of the eligible households participated in the housing allowance experiments even when to do so involved significant financial benefits. There is a serious question whether the integrity of such experiments has been undermined by such low rates of participation. Fourth, despite the fact that most of the large-scale experiments have been multisited, there is no small sample of sites in the United States that can be regarded as a fair sample of the conditions to be encountered in the country as a whole. Many researchers (Rossi and Wright, 1977) have urged the use of national samples rather than of a small number of sites, but the logistic problems of administering treatments to a dispersed national sample have generally been regarded as too severe to overcome at reasonable cost. Finally, almost all of the large-scale experiments have extended over very long periods of time. A reasonable rule of thumb is that it takes at least as long to analyze the data resulting from an experiment as it takes to collect the data involved. Thus the final results of the housing allowance demand experiment were available seven years after the start of the experiment and additional results will become available over the next few years. Clearly, large-scale true experiments are not the answer to urgent information needs.

Of course, these characteristics are ones that reign at this point in the development of the art of large-scale randomized field experiments, with which we have had experience for a little more than a decade. Further technical developments may make it possible to correct some

of the deficiencies, although it does appear that the time needed for such experiments will not be sufficiently shortened to make such research useful for short-term policy information needs.

Quasi Experiments: Making Do with Less than Perfect

The term "quasi experiment" is used to designate research that attempts to approximate randomized controlled experiments without the use of randomization to establish experimental and control groups. Approximations to equality between persons (or other units) exposed to a treatment and some sort of comparison persons (or other units) that were not so exposed is achieved through the use of nonrandomized controls, such factors as might conceivably affect the outcome being held constant through a variety of statistical techniques. Thus, for example, in the previously cited Skipper and Leonard (1968) study, a number of parents might have been asked to volunteer for special preoperative sessions with hospital personnel, the postoperative consequences of which would be assessed by comparing outcomes on such children with those on children whose mothers were not asked to participate, while such variables as age, sex, race, previous medical history, and other factors pertaining to the children that might be related to postoperative response are held constant.


Quasi experiments are sometimes set up prospectively by identifying an "experimental group" to which treatments are administered and a pre-existing "control" group that is not provided the treatment. Note that because randomization is not used in the definition of such "experimentals" and "controls," the two may differ in some significant way. Other quasi experiments may rely on retrospective data, asking individuals about their preselected exposure to particular sorts of treatments. For example, a quasi experiment on the effect of natural childbirth on perinatal health might ask a sample of recent mothers whether or not the natural childbirth procedures were used in labor, contrasting perinatal health conditions of children born under one condition with those born under other conditions, other known factors related to perinatal health status being held constant.

The utility of quasi-experimental designs, described in greater detail in Cook and Campbell (1979), lies in the considerably lesser costs

involved in time and data collection, as well as in the fact that they are often usable in circumstances where randomization is not feasible for one reason or another. The drawbacks of quasi-experimental designs have been fully described elsewhere (Cook and Campbell, 1979; Campbell and Stanley, 1966) and essentially boil down to whether or not adequate controls have been used in equating statistically those receiving the treatment with those who have not received the treatment. Whether or not a set of statistical controls is adequate depends on the fulfillment of two conditions: first, that at least the major contaminating variables be identified, and second, that reliable and valid operational measures of such variables be included in the study. The first condition depends upon a priori knowledge and theory. Such knowledge and theory are often either contradictory or simply non-existent. The second condition depends on the state of the art in the measurement of the variables in question, often a very primitive state indeed.

To illustrate these points we return again to our quasi-experimental version of a study of maternal stress and childrens' hospitalization. Rather than randomly grouping mothers into those who received extra attention at the hospitalization of their children and those who did not (as did Skipper and Leonard), suppose that volunteers were sought from among all mothers of children scheduled for tonsillectomies during a designated period, in order to accumulate an experimental group. Children whose mothers did not volunteer were then used as controls. As investigators we probably would have been wise enough to control statistically for mothers' education, but suppose "positive feelings about the hospital" was a potentially contaminating variable? Without including such a measure, we could not equate the two groups. But, even if we had been wise enough to consider the role of mothers' confidence in the hospital, would we have had the resources to develop a reliable and valid measure of this phenomenon?

Given the state of our current knowledge concerning the effects of all sorts of micro- and macrosocial factors on health, would-be experimenters cannot draw on sufficiently well-developed a priori theory that would identify all the potentially strong contaminating factors in the quasi-experimental version of the Skipper and Leonard experiment. One of the major virtues of randomization is that it is not necessary to know in advance what such factors might be, since randomizing effectively equates experimentals and controls on *all*



potentially contaminating factors, provided numbers are sufficiently large. Nonrandomly selected controls may or may not be adequate; the distressing fact is that, without randomization, we can never be sure.

Meeting the requirements for using various statistical controls is often equally dismaying, particularly to the purist in the evaluation field. The health area can learn from the controversy that still surrounds the now famous Westinghouse investigation of the impact of Head Start preschool programs on subsequent education performance. Fundamentally, the questions raised involved whether or not the failure to find differences between the "experimental" and the "control" groups was a function of lack of effect or a function of the inability either to meet statistical assumptions involved in the multiple regression procedures or because of the unreliability of measures, or both of the latter reasons. Clearly, findings about impact and lack of impact of interventions evaluated by quasi experiments are much more arguments of persuasion than discussions of experimental design philosophies. In general, not only are quasi experiments more open to controversy, but they frequently involve greater complexities in analyses, so that the "truth" derived from the simple elegance of the random experimental design is obscured.

But quasi experiments often are the only option. As we have noted, not only are some studies impossible to do otherwise, for one reason or another, but many true experiments have to be patched up through quasi-experimental analyses (e.g., Rossi, Berk, and Lenihan, 1980). Since almost all studies in the health field today involve a fair degree of "volunteerism" (subjects must elect to participate and to remain in the group assigned), sampling losses in the form of dropout rates are often so excessive—two or three times what would be regarded as significant statistical and/or policy differences—that controls have to be introduced statistically anyway. In many cases, patched-up experimental designs, because they generally have smaller sample sizes and limited measurement of contaminating effects, are less desirable than quasi experiments carefully planned from the outset.

Overcoming Field Intervention Limitations

The Guatemalan study of nutrition and cognitive development among very poor rural children in underdeveloped countries illustrates im-

portant common practical limitations in the design of field experimental studies (Freeman et al., 1977). On the basis of laboratory researches and loosely designed field studies, it seemed plausible to advance the hypothesis that malnutrition, even mild and moderate deficiencies, were linked to cognitive competence. With strong support from the National Institute of Child Health and Human Development, a series of parallel investigations were developed to test this basic hypothesis. One of the major efforts was undertaken at the Institute of Nutrition in Central America (INCAP).

An early research plan was to undertake a conventional epidemiological investigation in which young children would be measured on such indicators as height and head circumference, and the results correlated with performance on a battery of psychometric tests. Demonstrating causal relationships was impossible in such a study because of the difficulties, under these circumstances, of taking into account measures that would lead to the rejection of competing hypotheses about biological and social structural differences that were known to affect cognitive performance.

Undertaking a true experiment, however, was not feasible since the only way to intervene had to be on a village-wide basis, necessitating the inclusion of a relatively large number of villages as experimental units. The practical compromise was a design in which four "similar" villages were selected. Children living in those four villages were observed to age seven (as well as their mothers during pregnancies). Data on height, weight, and cognitive performance were gathered annually, providing the child and family were in the village at the time of the tests and examinations.

In addition, in two of the villages, a high-protein, high-calorie supplement was provided to the target populations. The original plan was simply to compare cognitive performance measures in the two villages given supplements with measures for the two unsupplemented villages.

The simple plan of comparing results of the two exposed and unexposed pairs of villages turned out to be overly simple. Not only did the children in the two supplemented villages vary in the extent of participation, but some of the children in the so-called control villages were adequately nourished. Accordingly, the basic analysis scheme was revised so that nutritional outcome measures (for example, height) were correlated with cognitive scores, and measures reflecting

various competing alternative explanations for cognitive performance were taken into account, particularly those related to family social class and social stimulation.

These regression analyses, as well as less powerful similar analyses in which quantities of supplementation consumed over exposure period were the independent or intervention variables, demonstrate that for many domains of cognition, nutritional status is a determinant of intellectual competence. The argument, however, is one of persuasion. The study's rigor suffers because it is impossible to dismiss two crucial criticisms: first, that only some of the contaminating influences have been controlled; and, second, that the measures of the influences eliminated are not sufficiently robust.

In addition, the study illustrates the "nesting problem"—the subjects are not independently selected for assignment to one group or another. In his review of evaluations, Cronbach (n.d.) has concluded that in educational research the classrooms, rather than the children, are most often the sampling units. The same issue needs to be dealt with in terms of many health services researches.

Notwithstanding the difficulties of this quasi experiment, it has yielded critical substantial data about an important basic research and social policy problem. Further, the sample size of more than 1,000 children at different time points and the rich data base of literally hundreds of variables allow a variety of different analyses and, as Cook and Campbell (1979) refer to it, considerable opportunity for "triangularization" to maximize the plausibility of impact or the lack of it.

Existing Programs and Policies

True experiments and quasi experiments are appropriate especially for testing out programs or treatments that are under consideration or possible modifications of existing programs that are under consideration but are difficult to apply to programs that have been in place and are essentially uniform in application over an entire political jurisdiction. In such cases it is difficult to identify control observations, persons or units who have not been exposed to the program and who are not obviously inappropriate for comparison with those who have received the intervention in question. Hence, to use a trite example, it is not possible to evaluate the effects of Old Age and Survivors Insurance (OASI) payments upon retired persons because those who

are of comparable age and not receiving payments either had not been working in covered employment categories, had not worked at all and were not married to covered workers, had refused to apply for payments, or were rendered ineligible for a small number of quite esoteric reasons. Any control groups made up of such persons or control observations made on uncovered persons are simply so contaminated with a variety of factors that it would be impossible to sort out the effects of OASI payments independent of such factors.

Fortunately, not all existing, well-established programs are like OASI payments. First, many are not uniform across all the jurisdictions in the country. Thus it is possible to compare different local versions of programs, versions that vary in coverage, eligibility requirements, and so on. Thus Cutright and Jaffe (1977) have estimated the impact upon fertility rates of federal support for family planning clinics by relating fertility differentials to the activity levels of the clinics, there being sufficient differences in effort and coverage among counties and aggregates of counties in the United States.

Second, programs start up and are discontinued, providing opportunities in the form of before and after comparisons. Thus Robertson (1980) estimated the effect of driver education on automobile accidents involving 16- and 17-year-old persons in Connecticut by observing the differences in numbers of accidents for this age group in jurisdictions that dropped driver education as compared with those that did not. Similarly Watson et al. (1980) compared measures of accident severity for motorcyclists in states that dropped helmet requirements in the late 1970s with measures in states that retained such requirements after relevant federal restrictions on aid to state highway funds were altered in 1974.

Third, some programs are only vaguely defined, considerable latitude in actual provisions being left to state and local jurisdictions. Thus while there has been considerable pressure for hospital providers to develop quality control measures in connection with care delivered under Medicare, it has been left to hospitals to develop specific mechanisms. An excellent survey (Gertman et al., 1979) collected data that described the utilization reviews employed in a sample of approximately 1,000 hospitals, and related such data to measures of hospital utilization under Medicare reimbursement provisions.

Another example of a study that takes advantage of program variations is one supported by the Center for Disease Control in cooperation with the University of North Carolina and the University of

California at Los Angeles. Currently in final analysis stage, the study attempts to assess the utility of recommendations for containing nosocomial (hospital-induced) infections followed to varying degrees by hospitals. Although a recognized problem throughout the history of hospitals, nosocomial infections have become of concern because, in addition to mortality and patient discomfort, such infections entail literally billions of dollars in excess hospital costs per year.

The recommended program includes having hospitals assign special staff to infection control, providing surveillance of activities and projects to minimize the transmission of infections. The design of this unusually ambitious study is unique. In over 350 hospitals, randomly selected to represent general hospitals in the United States, representative samples of patient records were reviewed for signs of nosocomial infections. In addition, a variety of persons in each of the hospitals, between 25 and 30, were interviewed in order to establish the character of the hospitals' infection program and the practices followed by medical and nursing staffs in them (Haley et al., 1980).

Not only is the study unusual in the size of its data set, but unique in its design. In addition to cross-sectional data in a given year on both program practices and incidence of nosocomial infections, the records before program establishment in each of the hospitals were checked in order to have available a baseline of "pretest" values for the dependent or outcome variable. Although the full analysis is not completed, the evidence that has emerged from analyses on relations between the program and hospital practices in patient care suggest at least some efficacy for the intervention package.

"Fine-Tuning" Efforts

In the introduction of this paper, the growing attention paid to increasing the efficiency of programs and health delivery systems was described. An analogy in clinical medicine is the discovery that a particular pharmaceutical intervention may be effective in managing most cases of a particular illness, but leaves a residue of unresponsive cases, as well as some instances of undesirable side effects.

In terms of access to care, this is the current state of affairs: The increased supply of physicians, stimulated by third-party payments (particularly governmental reimbursement), the development of new health practices, and the emergence of a network of neighborhood

or community health centers have all markedly reduced access differentials. Still, many persons in the United States, including rural and inner city residents, have insufficient access to health care compared with the rest of our population. Further, both formal studies and expert impressions offer reason to be concerned that some of the care received by these groups is of limited quality or is provided without sufficient regard to the dignity of patients.

Because of a strong commitment to equality of access, the Robert Wood Johnson Foundation has undertaken a number of large-scale demonstration programs to refine the delivery of ambulatory health services provided by neighborhood health centers. In one way or another, these fine-tuning efforts seek to link neighborhood and community health centers to the health delivery network by having them integrated and sponsored by medical schools, large public and community hospitals, and so on.

One of their major programs, the Community Hospital Program, provides support for establishing ambulatory group practices in, around, or as satellites of urban community hospitals. In order to measure access, a subsample of some twelve of these ambulatory group practices was selected and household interviews undertaken before and after project establishment. Andersen, Aday, and associates at the University of Chicago thus will have an opportunity to measure the differential use of health services as well as differences in the sites and in the providers of services before and after the activation of these hospital-linked sites for medical care.

In addition, a second group of investigators, Shortell and his associates at the University of Washington, are examining the emergence and changes in organizational arrangements of these health provider groups so that impact can be gauged at both the patient and the organizational levels. This study is still ongoing. Taken together with similar investigations of varying organizational linkages for providing health care to those with limited previous access, it is expected that a body of knowledge will accumulate, valuable in the face of continued and expanded governmental participation in the provision of health services (Aiken et al., forthcoming).

This access study, like the others described, is expensive and lengthy, requiring a pool of talented and committed investigators able to deal successfully with the complex technical problems and the many administrative and community relations challenges that are

encountered inevitably in these investigations. There are, of course, many quasi experiments of small size, including, for example, the already referenced spate of work to assess the impact of various deductible arrangements in health insurance programs on health behavior and medical costs.

At the same time, it should be emphasized that the sample sizes and number of control variables required in most quasi experiments, in order to have some reasonable confidence in the results, require relatively large-scale investigations. Further, the statistical procedures that allow for the most definitive inferences generally require technically sophisticated analysts and considerable computer resources. Thus, they are rarely of small magnitude, particularly for studies undertaken for policy-relevant and program purposes, as in the case of the last example. .

Monitoring Interventions. Although monitoring studies are an increasingly important aspect of program evaluation, they are not ordinarily considered to be an integral part of the experimental and quasi-experimental tradition. The purpose of monitoring studies is seen to be essentially descriptive, concerned with measuring the extent to which programs are reaching the subjects to which they are directed, the fidelity with which a program is being delivered, and the integrity of fiscal practices followed. Yet, we believe that monitoring activities will change in their concern from description to analysis, for reasons we will give below. Monitoring activities are included in this article for that reason.

In the human services field, when evaluation after evaluation indicated that programs more often failed to have any significant effects than to be successful in achieving their intended purposes, attention began to be given to an earlier question whether programs were being delivered as intended. After all, if a program is not being delivered, or is being delivered with changes that undermine its effectiveness, then it is no wonder that experimental or quasi-experimental evaluations arrive at the diagnosis of ineffectiveness. In human service after human service it was quickly found that program implementation was problematic; indeed, some programs were found not to exist at all after supposedly having been implemented; others were delivering treatments at such weak levels or in such transformed modes that the program could not be said to exist.

As a consequence, there is increased stress on monitoring evaluations. Although the health field may be no more defective than other human service areas of activity, many health interventions, both innovative and established, obviously fail to be undertaken as planned. This failure is often related to the very real difficulties with the problems encountered in operationalizing program elements and criteria for target population specification. Sometimes program staff appear to resist changes involved in particular programs. But implementation difficulties are often a consequence of political ideologies, and sometimes ethical considerations seemingly involved in the programs.

Although it may seem blatantly obvious that there is no point to studying the impact of programs that are not appropriately implemented in one way or another, without careful scrutiny of the intervention process it is not possible to know whether the program fails to impact as intended, or whether it is a matter of implementation. In the health services area, it is not easy to monitor most programs. First, the practice of medical care is rooted in the idea that "professionals" are responsible people, and it is regarded as insulting to question the performance of such professionals. This sensitivity of health professionals, of course, not only affects monitoring efforts but also day-to-day efforts at cost containment, maintaining quality of care, and increasing provider productivity. We may have come a long way from the sanctimonious position the physician had earlier in this century regarding privacy and autonomy, but cries of "interference" still persist when the turf of the provider is invaded.

A second aspect is that monitoring, at best, is inconvenient; at worst, it uses up time and resources that providers feel should be devoted to "practice." While there are some prospects for using unobtrusive measures in monitoring programs, for the most part they require observation, interviews, and additional record information.

At the same time, not only is the pressure for accountability by community members and resource support groups still on the increase, but also there is marked competition for available resources among apparently worthy programs. Take, for example, the area of new health practitioners discussed earlier in describing the Lewis and Resnick study. Even a decade ago the evaluation question was whether or not professional persons who were not physicians could render care

of reasonable quality in ways that would be accepted by patients. New health providers were seen as a solution to the then serious shortage of physician practitioners. Now, of course, although not many physicians are yet found on the welfare rolls, in many parts of the country there is clearly an adequate, or even an oversupply of doctors; consequently, comparisons of the ways doctors and near-doctors work have to be much more fine-grained than when the alternatives were a nurse practitioner or physician assistant or no one.

At first glance, it may appear simple to monitor programs. The delivery of health services is complicated, however, and often almost impossible to explicate. Monitoring studies are not simply a matter either of sitting around and observing, or of tabulating a few measures from existing records. If one is to undertake the monitoring of programs with a semblance of scientific rigor, then the effort almost always represents a major investment by program and evaluation staff. The importance and the complexities of monitoring evaluations are illustrated by two examples discussed below.

Schools as Health Care Sites

Among the options available for providing health services to children, indeed to entire families (Porter et al., 1976), is the public school. Not only are the populations "semicaptive," but in most communities schools are viewed positively as an institutional force, and thus a sensible entrypoint. Particularly in low-income areas with minimal provider resources, the idea of delivering primary care in school settings appears attractive.

As a prelude to a national school-health demonstration effort, a program of primary health care was developed for Chicago's Posin-Robbins school district. Nurse practitioners under the supervision of a physician preceptor are expected to provide care for emergencies and minor acute illnesses, as well as undertaking health examinations, referrals, and other preventive activities for more serious health problems. One major question is whether this type of school health program is utilized to an extent that can be thought of as "cost-beneficial." Another major question is whether the care provided is a supplement rather than a substitute for treatment provided by other health providers.

In order to monitor the program, it was first necessary to develop an "encounter" form for nurses to complete whenever a child was contacted, a survey questionnaire in order to learn about the backgrounds of children and their other sources of care, and a computer program that includes algorithms that make it possible to track treatment regimens and outcomes. In order to do the study properly, essentially a health dossier had to be compiled on each individual child, and to be kept constantly up to date.

The developed system is now in place in a four-state program of a similar type. Results of the Posin-Robbins investigation and preliminary findings from the four-state study raise important policy questions. In general, schools are viable sites for the provision of health services, and there is a fair degree of utilization by students. At the same time many, indeed almost all, of the encounters are essentially of a "band-aid and aspirin" variety. It thus is an issue whether or not the school health staff are operating as parent-surrogates, rather than as medical care providers. Also, there are important questions about costs. Although nurse practitioners receive substantially lower salaries than physicians, productivity may be so much lower, and the other trappings of the program so high in overhead costs, that the expenditures per encounter are as high as the costs of transporting a child to a board-certified pediatrician in an adjacent community. Although it is premature to render final judgments about the utility of the efforts being implemented, without a systematic and reasonably rigorous monitoring effort the apparent usefulness of offering primary health care in school sites might have been unequivocally accepted (Kaplan et al., 1979).

Feeling Good Is Bad

Preventive health education via the mass media has always seemed an attractive way of increasing community members' personal participation in health care. There is a long history of efforts that suggest that, while it is possible to increase factual knowledge via the mass media, it is difficult to modify and amplify actual behavior.

One persistent view is that communicators have failed to develop effective means of transmitting health messages, and that this is the reason for the lack of efficacy of mass media competence. Given the

apparent success of *Sesame Street* in reaching a mass audience, it seemed eminently sensible to use the same strategy for transmitting information to families about health practices and preventive behavior. *Feeling Good* was a *Sesame Street*-type of program designed to reach a large national audience and developed by the same team of creative TV producers, The Children's TV Workshop. A systematic study of *Feeling Good*, however, clearly provides evidence of its failure. The *Sesame Street* "style" did not work with adults and with a different message. It simply was not possible to develop a format and style of presentation that resulted in a sustained viewing and a regular audience for the program. It was not viewed by large numbers, nor was there enough of a persistent audience to merit its continuance. Thus it had only a short life span.

The two programs described are somewhat typical of various efforts at monitoring. Both point to the need in the provision of health services not to be overly sanguine about the effectiveness of various delivery systems. Health planners and providers apparently often are overly optimistic and overly enthusiastic about the ease of program implementation. Adequate monitoring is a counterforce to this bias.

Monitoring research is currently in a very primitive stage, concerned mainly with the accurate description of the coverage and implementation integrity of social programs, a stage similar to that of evaluation researches before the last two decades of development. As evidence accumulates concerning the difficulty of implementing on a mass scale human services programs of all types, attention will shift to a more analytic problem, namely, what are the conditions under which programs of given types can be successfully implemented on a mass scale? As such analytic questions come to be asked, monitoring research will shift to the use of true and quasi experiments. Treatments to be tested will center around different ways of implementing programs, the effects of varying incentive systems on delivery personnel, on alternative formulations of treatment that can best aid delivery efficiently, and in retaining treatment integrity.

Future Directions for Social Experimentation and Evaluation

Our review of social experimentation and evaluation necessarily has been selective. As in the past, much of future work will be directed

at examining relatively small, often mundane programmatic issues. But the importance of these efforts should not be disparaged. The health care enterprise represents a seemingly bottomless pit in terms of governmental costs of health services. And public expenditures for health in our country still provide only a proportion of total costs, with households and individuals bearing a large fraction, mitigated somewhat by insurance plans. Further, from the standpoint of both etiological research and health services studies, there remain virtually an infinite number of opportunities to undertake experiments and evaluation.

At the same time, particularly in terms of these "ordinary efforts," it is important to emphasize the need for rigor in implementing social research procedures, timeliness, and targeted dissemination of findings. We know that simply doing sound work is not enough, but that the potential usefulness of efforts requires developing a utilization strategy as well (Weiss and Bucuvalas, 1980).

Finally, from the standpoint of contributions both to social science knowledge and to major social change developments, our ideological and political climate provides exceptional opportunities for large-scale studies of national importance. Given the increased attention being given to the chronically ill and aged, the pressure to integrate more welfare and other human service efforts with health care, and the increasing competition for resources, there are opportunities for the social researcher to undertake major and challenging health evaluations. Persons committed to improving health services by systematic experimentation and evaluation have an unusual opportunity to demonstrate the cogency of their position. The issue, as always, is whether or not we have the creativity, techniques, and stamina to meaningfully contribute to improved health services and the health status of community members.

References

- Aday, L.A., Andersen, R., and Fleming, G.V. 1980. *Health Care in the U.S.: Equitable for Whom?* Beverly Hills, Calif.: Sage Publications.
- Aiken, L.H., Blendon, R.C., Freeman, H.E., and Rogers, D.E. 1980. Evaluating a Private Foundation's Health Program. *Evaluation and Program Planning* 3 (2):119-129.
-

- Bennett, C.A., and Lumsdaine, A.A., eds. 1975. *Evaluation and Experiment*. New York: Academic Press.
- Bernstein, I.N., and Freeman, H.E. 1975. *Academic and Entrepreneurial Research*. New York: Russell Sage Foundation.
- Bessman, A.N. 1974. Comparison of Medical Care in Nurse Clinician and Physician Clinics in Medical School Affiliated Hospitals. *Journal of Chronic Diseases* 27:115-125.
- Boruch, R.F., McSweeney, A.J., and Soderstrom, E.J. 1978. Randomized Field Experiments for Program Planning, Development, and Evaluation: An Illustrative Bibliography. *Evaluation Quarterly* 2:655-695.
- Brown, J.D., Brown, M.I., and Jones, F. 1979. Evaluation of a Nurse Practitioner-Staffed Preventive Medicine Program in a Fee-For-Service Multispecialty Clinic. *Preventive Medicine* 8:53-64.
- Burkett, G.L., Parken-Harris, M., Kuhn, J.C., and Escovitz, G.H. 1978. A Comparative Study of Physicians' and Nurses' Conceptions of the Role of the Nurse Practitioner. *American Journal of Public Health* 68:1090-1096.
- Campbell, D.T. 1969. Reforms as Experiments. *American Psychologist* 24:409-429.
- , and Stanley, J.C. 1966. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally.
- Chambers, L.W., Burke, M., Rosee, J., and Cantwell, R. 1978. Quantitative Assessment of the Quality of Medical Care Provided in Five Family Practices before and after Attachment of a Family Practice Nurse. *Canadian Medical Association Journal* 118:1060-1064.
- Connelly, S.V., and Connelly, P.A. 1979. Physicians' Patient Referrals to a Nurse Practitioner in a Primary Care Medical Clinic. *American Journal of Public Health* 69:73-75.
- Cook, T.D., and Campbell, D.T. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand-McNally.
- Cronbach, L. n.d. Kids in Classrooms. Unpublished manuscript.
- Cumming, E., and Cumming, J. 1957. *Closed Ranks*. Cambridge, Mass.: Harvard University Press.
- Cutright, P., and Jaffe, F.S. 1977. Estimating Family Planning Program Effects on U.S. Fertility Rates. *Evaluation Quarterly* 1:381-398.
- Dodd, S. 1934. *A Controlled Experiment on Rural Hygiene in Syria*. Beirut: Publications of the American University of Beirut, Social Science Series, No. 7.
- Ford, L.C., and Silver, H.K. 1967. The Expanded Role of the Nurse in Child Care. *Nursing Outlook* 15:43-45.
- Fox, R.C. 1959. *Experiment Perilous: Physicians and Patients Facing the Unknown*. Glencoe, Ill.: Free Press.

- Freeburg, L.C., Lave, J.R., Lave, L.B., and Leinhardt, S. 1979. *Research Report Series, Volumes I-IV*. Hyattsville, Va.: National Center for Health Services Research.
- Freeman, H.E. 1977. The Present Status of Evaluation Research. In Guttentag, M., ed., *Evaluation Studies Review Annual* 2:17-51. Beverly Hills, Calif.: Sage Publications.
- , Klein, R.E., Kagan, J., and Yarbrough, C. 1977. Relations between Nutrition and Cognition in Rural Guatemala. *American Journal of Public Health* 67:233-239.
- , and Solomon, M.A. 1979. The Next Decade in Evaluation Research. *Evaluation and Program Planning* 2:255-262.
- Garfield, S.R., Collen, M.F., Feldman, R., Soghikian, K., Richart, R.H., and Duncan, J.H. 1976. Evaluation of an Ambulatory Medical-Care Delivery System. *New England Journal of Medicine* 8:426-431.
- Gertman, P., Monheit, A.C., Anderson, J.J., Eagle, J.B., and Levenson, D.K. 1979. *Utilization Review in the United States: Results from a 1976-1977 National Survey of Hospitals*. *Medical Care* 17 (Supplement):1-103.
- Haley, R.W., Quade, D., Freeman, H.E., Bennett, J.V., and the CDC SENIC Planning Committee. 1980. Study on the Efficacy of Nosocomial Infection Control (SENIC Project): Summary of Study Design. *American Journal of Epidemiology* 111 (Special Issue):472-485.
- Kaplan, S.H., Berman, B.A., and Meeker, R.J. 1979. Evaluation of the National School Health Services Project. Presented at the 107th Annual Meeting of the American Public Health Association, New York, November 4-8.
- Levine, J.I., Orr, S.T., Sheatsley, D.W., Lohr, J.A., and Brodie, B.M. 1978. The Nurse Practitioner: Role, Physician Utilization, Patient Acceptance. *Nursing Research* 27:245-254.
- Lewis, C.E., and Resnick, B.A. 1967. Nurse Clinics and Progressive Ambulatory Patient Care. *New England Journal of Medicine* 277:1236-1241.
- Madge, J. 1962. *The Origins of Scientific Sociology*. New York: Free Press.
- McIntyre, C. 1894. The Importance of the Study of Medical Sociology. *Bulletin of the American Academy of Medicine* 1 (February):425-434.
- Muller, A. 1980. Evaluation of the Costs and Benefits of Motorcycle Helmet Laws. *American Journal of Public Health* 70:586-592.
- Newhouse, J.D., Rolph, J.E., Mori, B., and Murphy, M. 1980. The Effect of Deductibles on the Demand for Medical Care Services. *Journal of the American Statistical Association* 75:371.

- Pesznecker, B.L., and Draye, M.A. 1978. Family Nurse Practitioners in Primary Care: A Study of Practice and Patients. *American Journal of Public Health* 68:977-980.
- Porter, P.J., Leibel, R.L., Gilbert, C.K., and Fellows, J.A. 1976. Municipal Child Health Services: A Ten-Year Reorganization. *Pediatrics* 58:704-712.
- Riecken, H.W., and Boruch, R.F., eds. 1974. *Social Experimentation: A Method for Planning and Evaluating Social Intervention*. New York: Academic Press.
- Robert Wood Johnson Foundation. 1979. *Annual Report*. Princeton, N.J.: The Robert Wood Johnson Foundation.
- Robertson, L.S. 1980. Crash Involvement of Teenaged Drivers When Driver Education Is Eliminated from High School. *American Journal of Public Health* 70:599-603.
- Rosen, G. 1979. The Evolution of Social Medicine. In Freeman, H.E., Levine, S., and Reeder, L.G., eds., *Handbook of Medical Sociology* (third edition), 23-50. Englewood Cliffs, N.J.: Prentice-Hall.
- Rossi, P.H., Berk, R.A., and Lenihan, K. 1980. *Money, Work and Crime*. New York: Academic Press.
- , Freeman, H.E., and Wright, S.R. 1979. *Evaluation: A Systematic Approach*. Beverly Hills, Calif.: Sage Publications.
- , and Wright, S.R. 1977. Evaluation Research: An Assessment of Theory, Practice, and Politics. *Evaluation Quarterly* 1:5-52.
- Shortell, S.M., and Richardson, W.C. 1978. *Health Program Evaluation*. St. Louis: Mosby.
- Simborg, D.W., Starfield, B.H., and Horn, S.D. 1978. Physicians and Non-Physician Health Practitioners: The Characteristics of Their Practices and Their Relationships. *American Journal of Public Health* 68:44-48.
- Skipper, J.K., Jr., and Leonard, R.C. 1968. Children, Stress and Hospitalization: A Field Experiment. *Journal of Health and Social Behavior* 9:275-287.
- Spector, R., McGarath, P., Alpert, J., Cohen, P., and Aikens, H. 1975. Medical Care by Nurses in an Internal Medicine Clinic: Analysis of Quality and Its Cost. *Journal of the American Medical Association* 232:1234-1237.
- Spitzer, W.O., Roberts, R.S., and Delmore, T. 1976a. Nurse Practitioners in Primary Care. V. Development of the Utilization and Financial Index to Measure Effects of Their Deployment. *Canadian Medical Association Journal* 114:1099-1102.
- , Roberts, R.S., and Delmore, T. 1976b. Nurse Practitioners in Primary Care. VI. Assessment of Their Deployment with the

- Utilization and Financial Index. *Canadian Medical Association Journal* 114:1103-1107.
- , Sackett, D.L., Sibley, J.C., Roberts, R.S., Gent, M., Kergin, D.J., Hackett, B.C., and Olynich, A. 1974. The Burlington Randomized Trial of the Nurse Practitioner. *New England Journal of Medicine* 290:251-256.
- Suchman, E. 1967. *Evaluative Research*. New York: Russell Sage Foundation.
- Sullivan, J.A., Dachelet, C.Z., Sultz, H.A., Hency, M., and Carrol, H.D. 1978. Overcoming Barriers to the Employment and Utilization of the Nurse Practitioner. *American Journal of Public Health* 68:1097-1103.
- Watson, G.S., Zador, P.L., and Wilks, A. 1980. The Repeal of Helmet Use Laws and Increased Motorcyclist Mortality in the United States, 1975-1978. *American Journal of Public Health* 70:579-585.
- Weiss, C.M., and Bucuvalas, M.J. 1980. Truth Tests and Utility Tests: Decision-Makers' Frames of Reference for Social Science Research. *American Sociological Review* 45:302-312.

Address correspondence to: Peter H. Rossi, Director, Social and Demographic Research Institute, W-34 Machmer Hall, Amherst, MA 01003.