Educational Evaluation and Policy Analysis
Fall, Vol. 4, No. 3, 1982, pp. 261-279

Ibsen's Baths: Reactivity and Insensitivity

(A Misapplication of the Treatment-Control Design in a National Evaluation)

David M. Fetterman
RMC Research Corporation and Stanford University

This discussion examines the insensitivity of the design and conduct of a national evaluation study. The fundamental problem with the evaluation is that it typifies a misapplication of the treatment-control or experimental design in educational research. First, the application of this design to a population of dropouts is unintentionally unethical, denying dropouts the opportunity to come back into the social system by entering an alternative school. Second, there are significant methodological flaws in the application of this design. The control group in this study is actually a negative treatment group or a reactive control group (Cook & Campbell 1979) composed of individuals refused participation in the program. The claim that withholding treatments from a portion of the group will provide more accurate estimates of the treatment effect (as argued by Tukey, 1978, and Gilbert, McPeek, & Mosteller, 1978) is spurious in this case. Controlled trials can actually ensure a more inaccurate estimate of treatment effects. This discussion demonstrates how a control group does not provide the "no-cause" baseline information expected of it.

A discussion of this fundamental paradigm in educational evaluation parallels the controversy surrounding Ibsen's baths in his play *An Enemy of the People* (1958). Dr. Stockmann, the play's protagonist, was a medical official of the town's baths. After discovering contaminants in the baths, Stockmann attempted to publish his findings. He encountered significant resistance from the townspeople, who derived their income from tourists frequenting the baths. I have encountered similar resistance to the discussion of the misuse of a paradigm fundamental to educational research. Dr. Stockmann believed that "The whole of [our] flourishing municipal life derive[d] its sustenance from a lie" (p. 627). The author contends that a parallel exists in an evaluation that relies on information drawn from a misapplication of the treatment-control design. Stockmann's community and the evaluation community are blinded by different agendas. The difference between the two is that Stockmann's findings centered around personal and economic motivational factors, while the misapplication of the experimental design in this case was the result of real world constraints and insensitivity. The results, however, are the same: both forms of blindness and resistance to substantive criticism function at the expense of others (in each case the individuals they seek to assist). A brief examination of the real world constraints and views of the federal bureaucracy, the research corporation, and the education research establishment

I am indebted to Lee J. Cronbach and my colleagues at RMC Research Corporation and Stanford University for their insightful criticism and generous assistance in the preparation of this manuscript.

illuminates the reasons why policymakers and researchers continue using this design.

Context

Congress approved the Youth Employment Demonstration Project Act (YEDPA) legislation in 1977. The act assigned $1.6 billion to manpower programs for youth and there were approximately $110 million to be spent in discretionary demonstration efforts that would foster "knowledge development" in the area of training and employment of "disadvantaged youth."

The author is currently responsible for the ethnographic component of a national evaluation. One of the tasks included reviewing a large sample of these demonstration efforts. Informal interviews were conducted with over 50 program directors across the country, with particular emphasis on their evaluation designs. One demonstration project evaluation has been identified that typifies the experience of many programs. This demonstration project is based on an exemplary school for unemployed high school dropouts. The objective of the program is to provide dropouts with a career orientation and a high school diploma. The aim of the project is to replicate the original model of the exemplary school in four different parts of the United States.

A federal agency developed a proposal to test whether this program could be set up in different communities and to show its original effectiveness with the youths admitted to the program—to qualify as a demonstration project. The request for proposal (RFP) required the four new replicating sites to repeat the evaluation design used in the prototype site.

The original design had both psychometric and ethnographic components to answer different kinds of questions. The psychometric component of the evaluation (basically adopted by the current evaluation and the focus of this discussion) is described in the RFP.

> The design involved three separate cohorts of applicants, applying at six- to eight-month intervals. Through over recruitment and a lottery process, known in advance to all applicants, three separate sets of experimental and control groups were selected in a ratio of about three students to one control. The ratio was selected to permit maximum entry into (the program) with the minimum N estimated to be needed for a group large enough to be sensitive to educationally meaningful effects. (Reference deleted)

A program staff member's remarks concerning this evaluation design echoed the sentiments of the disseminators, adopters, a few LEA (local educational agency) members, and community members (at the sites): "How many times are we going to be used as guinea pigs. . . . We have real problems that need help now, not more demonstration projects. . . . They already proved it was an 'exemplary' program. . . . Why aren't they just trying to help us make it work." There are a variety of concerns expressed in this individual's remarks including disillusionment with demonstration projects that leave the community as quickly as they arrive, alarm and resentment concerning the experimental terminology[1] of the evaluation, and most pertinent to this discussion, outrage regarding the ramifications of the treatment-control design itself.

It is difficult to convey the importance of additional trials to individuals who witness the day-to-day deleterious effects of those trials on "disadvantaged" participants. The experience of Follow Through, however, exemplified how "commitment may precede adequate planning and developmental inquiry" (Elmore, 1975; Haney, 1977), as well as the elusiveness of a perfect social experiment. The application of this design demonstrates a desire to solve these social problems and to respond to accountability concerns in government (e.g., efficient use of taxpayer money). The absence of a rigorous evaluation represents an abdication of responsibility; however, a rigorous evaluation need not always employ the treatment-control design. Moreover, the repetition of this design in

[1] A number of individuals associated the "experiment" with experiments conducted with rats because of the treatment-control terminology. Potential students would periodically ask if the tests were designed to "tell 'em if I were crazy or somethin'." Although the testers assured them that this was not the intended purpose, many students left the test sessions still under the impression that there was a "hidden purpose."

new sites throughout the country is not a useful or appropriate mechanism for establishing external validity, contrary to Cook and Campbell (1979) and the GAO guidelines.

A Misapplication of the Treatment-Control Design: Ethical, Programmatic, and Methodological Problems

First, the application of this design to dropouts and students whom the urban high schools are currently failing to serve (potential dropouts) is unintentionally unethical. In the federal bureaucracy's desire to present conclusive findings based on a rigorous research design—to present "absolute scientific proof" of the program's success or failure—it failed to recognize the ethical consequences of applying this design to primarily low-income, minority, high school dropouts and potential dropouts. Briefly, the problem is that human beings are being denied a second chance, for many their last chance, to function productively within the system. Teenagers, primarily low-income, minority students (or ex-students) who were disenchanted with "the system" and dropped out or were about to drop out, decide to return to the school setting (a symbol of the larger sociocultural system, and a formal mechanism of socialization). Through the lottery process, however, at least one-fourth of them are told to look elsewhere for assistance. The effects of turning away an individual whose desire to lead a productive life has been "rekindled" are numerous and profound. The tears of a young woman (heard over a transcontinental call) who has received a letter assigning her to the control group are only a small part of the problem. Interviews with rejected students and their parents at each of the sites reflect similar concerns regarding their child's pattern of behavior; they are "falling back into their old ways, not goin' to school, not working, just hanging around with so and so, those hoodlums."

This is an old argument recently revived in clinical work by Tukey (1978) and Gilbert et al. (1978) in *Evaluation Studies Review Annual*. Cook and Campbell (1979) and Boruch (1976) have also written extensively about the issue in the context of evaluation studies. The basic argument has revolved around the issue of costs: the cost to a few denied "treatment" as juxtaposed to the costs to the vast majority that suffer from ambiguous results when the experimental design is not applied to the investigation. Tukey's (1978) position in this argument is clearly presented:

> The pressures of ethics and equity on clinical trials have always been severe. Today they are more vigorous than ever before. Many of us are convinced, by what seems to me to be very strong evidence, that the only source of reliable evidence about the usefulness of almost any sort of therapy or surgical intervention is that obtained from well planned and carefully conducted randomized and, where possible, double blind clinical trials [see the review papers of Byar et al. (1976) and Peto et al. (1977)]. Dare we prevent ourselves from obtaining reliable evidence? (p. 327)

Gilbert et al. (1978), who agree with Tukey, point out that two arguments emerge from this discussion and in each case "the patients are seen to be losers:"

> The first argument is an expression of the fear that the trial, by withholding a favorable new therapy imposes a sacrifice on the part of some of the patients (the control group). The second argument raises the opposite concern that, by getting an untested new therapy, some patients (those in the experimental group) are exposed to additional risk . . both arguments imply that investigators know in advance which is the favorable treatment. (p. 333)

The Gilbert et al., response to these arguments argues that the benefits to society—of accumulated knowledge through randomized controls—far outweigh the costs to the individual.

> If participation seems to the patient to be a sacrifice, it should be noted that others are making similar sacrifices in aid of the patient's future illnesses. So even if the particular (controlled) trial may not help the patient much, the whole system is being upgraded for his or her benefit. We have a special sort of statistical morality and exchange that needs appreciation. (p. 337)

The notion of relative sacrifice comes immediately to mind when comparing the sacrifice of a middle or upper middle class student deprived of a special program as juxtaposed to the sacrifice of a lower income dropout who is deprived of a second chance to enter the mainstream sociocultural system. In the former case the individual is generally provided with equally or near equally productive alternatives within the mainstream of the system. The sacrifice is more poignant in the latter case. The statistical morality presented by Tukey and Gilbert et al. is misleading in this case. This evaluation demonstrates how the social costs of denying treatment to an individual with few alternatives to enter the legitimate social system far outweigh the benefits derived from the application of this design and is, therefore, unethical on a much larger scale.

The comparison between medical and educational innovations is in itself misleading. The two units are not comparable. As Cronbach and Associates (1980) explain:

> For a technique of health care, a trial under highly controlled conditions almost always preceded a more realistic field test. An educational or welfare service is much less likely to be installed as a superrealization. The theory underlying the proposal is not so definite as the biological hypotheses from which a vaccine is derived, hence there is less interest in what is ideally possible. Also, where biologists have a tradition of patient step by step science, social reformers are impatient. Action moves ahead despite large uncertainties. (pp. 239–240).

Serious ethical and methodological problems also arise when the statistical morality logic is applied to the replication or dissemination of exemplary educational demonstration programs. (The methodological problems resulting from the replication effort are discussed in detail under threats to external validity; ethical considerations in this regard are discussed below. There is of course considerable overlap.) Exemplary programs by definition have already been proven successful as programs per se. There is good reason to believe that the program will be successful in a different geographic area with a similar population, if properly implemented. Individuals in this case are consciously being deprived (randomly) of a benefit. Boruch (1976) points out that "if it (the treatment) is known to be beneficial, then the experiment may well be unethical" (p. 187). The aim in a replication project is not to detect program effects per se but to duplicate prototypical program effects. When the means of an objective (detecting effects to establish if they have been duplicated) are confused with the ends (establishing if the program has been duplicated) serious problems arise. An undue emphasis is given to the significance of those findings and their role in an evaluation. Methodologically the problem is evident. Ethically, however, the problem is less evident. The evaluator is personally as well as professionally obligated to present the truth (or the best approximation of it). A recognizable methodological flaw seriously compromises the credibility of the research findings and the researcher. Consequently, presentation of unintentionally distorted findings represents a serious ethical dilemma. Subjecting a developing demonstration program, with all the implementation difficulties endemic to such an endeavor, to the effects of a treatment-control design and then comparing the results of replicating sites with those of more mature prototypes appears unconscionable. Methodologically, the demonstration programs' results are more likely to represent their implementation difficulties than a given educational program treatment. Researchers face a serious ethical dilemma when delivering this kind of comparison to policymakers, as discussed above. The fate of a program or a "beneficial" treatment is seriously jeopardized by such an unfair comparison, even though the comparison is mandated as part of the replication or evaluation process.

Boruch's (1975) tour de force "On Common Contentions About Randomized Field Experiments" is a useful guide in this endeavor. The section regarding techniques used to "reduce conflicts between ethical standards and evaluation needs," however, does not adequately respond to the constraints operating on the type of program under discussion. For example, in a 2-year demonstration secondary education

program where a year follow-up on graduates is required, "delaying treatment for individuals in the randomized control group" is unrealistic if not impossible. Treating members of the control group before the experiment and graduate follow-up has been completed contaminates results derived from the controls. Attempting to treat the control-group individual after the treatment group has completed its participation in the program is also problematic: participants become too old (and are therefore ineligible for the program or have pressing familial obligations); the majority are not interested in waiting around for the treatment; and/or the demonstration period has ended and the program may be closed. (Some of the limitations of this approach, as Boruch suggests, can be found in Chapter IV of Riecken et al., 1974).

Similarly, "playing the winner" is inappropriate for an educational demonstration project. Boruch (1976) explains how this strategy works:

> Here, subgroups of candidates or individuals are assigned to a program only as long as the outcome of treatment is successful. When a failure occurs, the very next subgroup or individual is assigned to the control (or alternative treatment) condition. (p. 188)

This strategy is only effective when "success or failure becomes evident quickly and when switches can be accomplished easily." Neither of these conditions exist in developing demonstration programs. Success or failure in demonstration programs is rarely evident quickly—to participant or to researcher—and switches often disrupt the continuity required in educational settings to effect significant change. There are similar difficulties with various other strategies including comparing treatment variations, assessment of components of a program and so on.

The fundamental argument, however, is that this application of the treatment-control design to this program is methodologically flawed because the claims for withholding treatments from some individuals in this study are spurious. Depriving certain (randomly selected) individuals from a "beneficial" treatment does not necessarily lead to more accurate estimates, as claimed by Tukey, Gilbert et al., and Boruch among others. The study under discussion illustrates how this randomized control claim and the statistical morality logic are spurious when control groups do not provide the "no-cause baseline" information required to complete the task. This is discussed in detail following a brief presentation of the exacerbating role of insensitivity in evaluation.

Insensitivity in Evaluation

The assignment of an individual to a control group is often exacerbated by the entire process by which the individual is turned away. The individuals, in this case, must first decide to give themselves or the system another chance. Then they must undergo approximately 5 hours of psychometric tests, pass the reading test (at a fifth-grade level), and are then thrown into a lottery system where all of them are exposed to the possibility of nonacceptance or failure again. It is important to emphasize that this program represents their "last chance," according to many of the potential students and their parents. Exposing these students to another opportunity to fail can inflict serious personal damage, as numerous interviews with rejected students suggest. Moreover, it is argued that the application of this design to this particular population (dropouts) produces unintentionally deleterious effects in two additional areas. First, it generates serious programmatic problems for the demonstration site. These problems are generally interpreted negatively by the evaluators and serve to misrepresent the program. Second, the application of this design to a dropout population raises serious questions about the credibility of this segment of the evaluation or any evaluation that uses this design with similar populations.

Programmatic Problems

The most serious programmatic problems generated by this design concerned recruitment. Program staff members faced an uphill battle to sell a program to potential students and their parents who considered it a risk: because it was a demonstra-

tion project, it was perceived by many as "a school for dropouts," and they needed to pass a test to gain admission to the program. When program staff members (or students who helped in recruitment) added that admission was further predicated on the youths' luck in being chosen by lottery to establish a control group, the appeal of the program was more than "just slightly tarnished." Basically, this design turned a number of students off before they even began to understand the program—erecting almost impenetrable barriers to recruitment, in the early period of the demonstration. (Subsequently, it erected similar barriers for alienated youth interested in working within "the system."). This was compounded by the community's negative perception of the program. One director of another alternative program explained: "What kind of program asks for kids so they can turn one quarter of them away, back into the streets. Why should we recommend that kind of program?"

The implementation of this design also had another unintentionally negative consequence for program operations, specifically for recruitment. The evaluation's impact on the program is directly related to the nature of the research corporation itself.

The evaluation corporation is a business concerned with producing a reputable research product, advancing the state of the art, and making a profit. Regarding the latter concern, professional testers were hired for this project in a manner intended to maximize their efforts (and minimize their costs). The professional testers were initially instructed not to test potential students unless 15 or more students could be identified for testing at the sites. Consequently, staff members were unable to inform many students when testing would occur. Students waiting for about 4 weeks lost interest in the program. The attrition rates between the initial interview and testing ranged from 26 to 59 percent.

Similarly, a second cohort was postponed because the recruitment figure was below the expected number (at least 75 youths). The evaluation design required that students enter as a block or cohort. Youths (already tested) were held in limbo for periods ranging from between 1 and 14 weeks for information regarding program initiation. Once again many lost interest in the program and found other avenues of interest to pursue. One site lost 49 percent and two other sites lost approximately 20 percent each of their potential students due to the waiting period between testing and intake. (Recommendations of an Advisory Panel were taken into consideration for a third cohort concerning this matter. Testing was conducted on a demand basis. This represented a more expensive procedure, but it reduced the attrition between testing and intake ranging from 7 to 15 percent). A recruitment barrier, however, remained throughout the demonstration project. An identifiable faction of students continued to resist program recruitment efforts; according to interviews, they feared the potential consequences of the lottery and testing obstacles—a sense of personal failure.

Methodological Problems

There are serious methodological flaws in the application of this design to a dropout population. In a classical treatment-control design (in educational and psychological experimentation) individuals are randomly assigned to treatment and control groups. The control group determines what the treatment or experimental group would be like without the special treatment. One of the classical paradigms involving human beings, for example, is drawn from pharmaceutical studies of drug effects. A segment of the population is randomly assigned to a treatment and control group. The treatment group is given the drug or treatment and the control group is given a placebo. (See Matarazzo, 1965, for elaboration of the "placebo effect" in educational and psychological evaluation.) Neither group is aware of who has been given the treatment. Knowledge of group membership (treatment or control) might significantly affect study outcomes. In the national evaluation under discussion, one segment of a biased sample is placed in the treatment group (individuals who were not "turned off" by the rigorous examination and lottery system) and another segment according to parents is "slapped in the face"—told they could not enter the program. This second group

represents the control group. In this case there is no control group, but merely a negative treatment group or a reactive control group[2] to be compared to a biased sample of "treatment" students (see Tallmadge, 1979, for elaboration concerning the problems with treatment and control groups as applied to specific programs and populations). Campbell and Stanley (1963) refer to this problem under reactive arrangements: "The reactive effect can be expected whenever the testing process is itself a stimulus to change rather than a passive record of behavior" (p. 9). Cook and Campbell's (1979) presentation of this problem is the most applicable to the study under discussion. The issue is found in the context of threats to internal validity under the heading "Resentful Demoralization of Respondents Receiving Less Desirable Treatments:"

> When an experiment is obtrusive, the reaction of a no-treatment control group or groups receiving less desirable treatments can be associated with resentment and demoralization, as well as with compensatory rivalry. This is because persons in the less desirable treatment groups are often relatively deprived when compared to others.... In an industrial setting the persons experiencing the less desirable treatments might retaliate by lowering productivity and company profits, while in an educational setting, teachers or students could "lose heart" or become angry and "act up." Any of these forces could lead to a posttest difference between treatment and no-treatment groups, and it would be quite wrong to attribute the difference to the planned treatment. Cause would not be from the planned cause, A, given to a treatment group. Rather, it would be from the inadvertent resentful demoralization experienced by the non-treatment controls. (p. 55)

The treatment sample is further biased by the type of students required to meet the demonstration project timelines. Only students seeking a diploma and with sufficient credits to graduate from the program within the allotted demonstration period were initially accepted into the program (11th and 12th graders). This type of student is often referred to as "the cream of the crop" by local high school personnel. One administrator of a feeder high school said that he believed many of the students would have made it anyway. The author disagrees with this judgment. A more accurate assessment of these students would be: they are the best of the lowest achieving students in the public schools, for example, bored, cutting classes, but still interested in graduating and with few serious legal or psychological problems. There are some indications that this administrator's comments represent a response to a perceived threat: if the program is successful with the students his school failed to serve, it would not reflect well on his school. The students, however, are clearly not the most disillusioned within the entire system.

There are also serious threats to external validity in this study, and it is argued that external validity or generalizability is the most important element of an evaluation. While it is a useful research question whether a functioning program in one city can function similarly in another city, the methodology employed to test this proposition in this study is questionable. Cook and Campbell (1979) "stress that external validity is a matter of replication" (p. 78). Similarly, GAO guidelines recommend the repetition of the experimental design in new sites to secure external validity. "The primary tool for establishing external validity is replication of the evaluation in diverse settings" (Comptroller General, 1978, p. 15) The evaluation design in this study followed GAO guidelines in this respect. Comparing and interpreting experimental outcomes, for example, graduation

[2] The problem was further compounded by the Hawthorne (Roethlisberger & Dickson, 1939) and John Henry effect (Saretsky, 1972). Regarding the Hawthorne effect, students entering the program perceived the process as winning in a competition, as evidenced by one student's comment: "I was one of the one's that got in! We're the best!" This perception produces another treatment separate from the so-called program treatment. In addition, there is some evidence (after a secondary analysis of the data) that a consistent block of control students display the John Henry effect. Interviews with some of these students indicated that they perceived the posttest experience as an opportunity to show the program personnel that they were wrong or that they were "as good as anybody." This problem is then further compounded by the problems of differential attrition in both treatment and control groups.

rates and reading scores, led to spurious conclusions in the absence of a baseline that represented the graduation rates and reading scores of that population in those cities at that time.

Furthermore, there is no reason to expect a program to replicate itself in a new site. The concept of replication is a biological, not a sociological or anthropological, concept. Sociological systems do not contain genes and thus do not follow the pattern of biological reproduction. This is precisely where the analogy between biological and sociological evolution breaks down (Fetterman, 1981). A program adapts to a new environment. Applying the same evaluation design to variations of a model will produce systematically different results. Cronbach and Associates (1980) take a strong stand on this position:

> Users are wrong to assume that what worked in the trial will work in the future. The treatment *T* will be changed into a *T** when it is applied on a larger scale or under other administrative arrangements; moreover, the PSC (policy-shaping community) will often want to depart from the plan originally tried. Transfer to a new population and setting will also modify results. In view of the fact that interactions abound, change in conditions or procedures can enhance, reduce, or even reverse the effect of a treatment (Campbell, 1974; Cook & Campbell, 1979, pp. 28, 33–35; Cronbach, 1975; Bronfenbrenner, 1976; Pillemer & Light, 1979). The only way the PSC can exercise judgment about future programs bearing the same label as the *T* studied is to understand the process by which the treatment works. Understanding is required to make use of even a well-grounded formal conclusion. (p. 275)

Similarly, Cronbach and Associates (1980) explain what can be learned from the range of variation that should be expected in the "replication" process.

> For purposes of a prototype study, treatment plan *T* may be installed in a number of sites. Even so, the realizations are almost certain to vary, if only because the guidelines are interpreted locally. In the Follow Through comparison of "models" of compensatory education the variation in results across sites supposedly using the same model was ten times the variation across models (Glass, 1979). Allowing natural variation to occur and then appraising its extent makes interpretation comparatively easy. If findings are consistent from site to site, the PSC learns that the treatment has much the same consequences wherever and however it is installed. Insofar as the results differ, something much more important is learned: not all realizations that come under the same general label work the same way, and a plan to establish a uniform program by a centralized decision may be a fantasy (Anderson and others, 1978). A more refined analysis will then become the basis for future planning. If nothing else, a close look at the less successful realizations can suggest guidelines that will make such deviations infrequent. But the findings may suggest establishing radically diffferent programs in settings that differ. (Pillemer & Light, 1979, pp. 277–278)

This position represents a better way to handle the problems discussed earlier regarding the national replication study. Moreover, the author is inclined to agree with Cronbach and Associates that "multiple, diversified, decentralized studies" will produce the most useful research for policymakers.

The author has carefully delineated when control groups do not provide the "no-cause" baseline required in an experimental design. In addition, threats to external validity have been briefly discussed. Paralleling Cronbach and Associates' (1980) position, the author has not presented an argument against controlled assignment per se, "but against the unsophisticated conception of random assignment as a magic bullet that kills off all threats to validity" (p. 304). The question remains, however, why the misuse of this design is not uncommon.

Real World Constraints and Perspectives: The Federal Bureaucracy, the Research Corporation, and the Educational Research Establishment

The key to understanding how this design can be misused repeatedly lies in the powerful role played by the real world constraints and views of: the federal bureaucracy, the research corporation, and the educational research establishment.

This analysis provides a rationale or logic for the events described throughout this discussion.

The federal bureaucratic perspective contributes to the misuse of the treatment control design, due primarily to environmental pressures. A brief examination of the federal agencies' real world constraints and views provides a rationale for the misuse of the research design. The perspective of federal government policymakers is clearly presented in the literature by Mulhauser (1975); Coward (1976); Holcomb, (1974); Etzioni, (1971); von Neuman and Morganstern (1953); March and Olsen, (1976); Acland, (1979); Cronbach and Associates (1980); Rich (in Weiss, 1977); Elisburg (1977); Lindblom and Cohen (1979); Baker (1975); among others. One of the primary responsibilities of the federal sponsor is to produce the most credible and socially relevant research (Holcomb, 1974) dictated by Congressional mandate. Policy research, in contrast to basic research, however, represents another significant facet of the federal bureaucratic perspective.

> [Policy research in juxtaposition to basic research] is much less abstract, much more closely tied to particular actions to be undertaken or avoided. While basic [research] aims chiefly to uncover truth, policy research seeks to aid in the solution of fundamental problems and in the advancement of major programs. (Etzioni, 1971, p. 8)

Policy research seeks immediate action in response to a troubled situation such as unemployment, a high dropout rate, and so on. It attacks a discrete facet of that situation to "avoid turf problems." Decisions are made in a context of accommodation rather than command (von Neuman & Morganstern, 1953). Policy is more a process of drifting toward a decision, than a Platonic pattern of a single commander handing down decisions affecting the entire social sphere (see March & Olsen, 1976). There is, according to Mulhauser (1975), "no search for a comprehensive understanding of the problem's nature or origin" (p. 311). Glennan (1972) pointed out that significant go/no-go decisions are rare in policy. Cronbach and Associates (1980) add to the picture the fact that:

> Policy makers do weigh alternatives that have incommensurable outcomes—reduced-crime-versus-community harmony, say, or children's-shoes-versus-Army boots. (p. 287)

There is simply a time pressure that requires immediate identification of politically viable "levers of action." Often, Mulhauser points out: "The action taken is a minor variation on what was done the last time something like this came up" (p. 311).

Federal agencies are also constrained by the responsibility for providing timely input for policymakers. As Coward (1976) points out, "Evaluation data presented after a policy decision has been made can have little impact on the decision" (p. 14). The role of evaluation itself is limited in the policy arena. It is used, according to Rich (in Weiss, 1977, p. 200), in "groups and clusters" as one piece of evidence or data in the larger, fundamentally political equation (Acland, 1979). Cronbach and Associates (1980) point out that "What impresses a research expert obsessed with method may not impress someone who sees the larger picture" (p. 294). Elisburg (1977), similarly places the Congressional role of evaluation into perspective:

> It cannot be stressed too strenuously that scientific program evaluation is itself evaluated by the Congress in terms of its utility to promote the effectiveness and precision of legislative judgements in a political milieu. (pp. 67–68)

Furthermore, according to Cronbach and Associates (1980),

> Knowing this week's score does not tell the coach how to prepare for next week's game. The information that an intervention had satisfactory or unsatisfactory outcomes is of little use by itself; users of the study need to know what led to success or failure. Only with that information can the conditions that worked be replicated, or modified sufficiently in the next trial to get better results. (p. 251).

In addition, federal agencies must maximize their returns in efforts with limited fiscal resources. Combining scarce resources with pressures of accountability

produces a climate of interagency rivalry over those resources and thus the need to employ the maximization model (McClelland & Winter, 1969). The maximization model suggests "that human beings everywhere tend to choose the personal action that they feel will gain them the greatest benefit (or avoid the greatest loss) with the smallest expenditure of resources" (p. 21). (See Bailey, 1960; Barth, 1963, 1966, 1967; Erasmus, 1961; Kunkel, 1970.)

These fundamental constraints shape the agencies' perspective and enable them to adapt successfully to the federal environment. The federal agencies' survival literally depends on an adequate understanding of, adherence to, and manipulation of these norms. The fluidity of funding from year to year, political fluctuations and alliances, career-building concerns, and the acquisition-maintenance of power games all contribute to the political instability of the bureaucratic hierarchy and federal perspective. "The political process has a life style and morality of its own—a lifestyle and morality that evaluators have to respect if they are to be of use" (Lindblom & Cohen, 1979, as paraphrased by Cronbach & Associates, 1980, p. 349).

The demands for data, according to strict guidelines and timetables, are generated from this environment. Knowledge is power, and information is required at prespecified periods to assist in the federal decision-making process—for example, assessing the relative merits of competing programs. Coward (1976) warns, "Agencies place themselves in highly vulnerable positions if they sponsor a research effort that is unable to provide data under constraints imposed by policy deadlines" (p. 14). The inability to address these concerns in this fashion may leave an agency "out in the cold," with little or no future funding. These constraints and the socialization of federal bureaucrats according to the canons of the traditional educational establishment (discussed later) have guided the federal government into the pattern of traditionally associating the most credible and timely research with the experimental design; regardless of the task at hand.

The description of the research corporation's effect on program recruitment demonstrated how its perspective influenced behavior. The corporation as a business was interested in maximizing their efforts and minimizing their costs. This orientation motivated the corporation to instruct their testers to test no less than 15 students at a time. The result was that students lost interest in the program while waiting for a large enough group to be assembled. The programmatic interference was unintentional. The disruption was simply the logical result of a research corporation's businesslike perspective.

Another facet of the research corporation's perspective is related to the misuse of the experiment paradigm: the pattern of bidding for proposals with the problem, and in many instances the research design, defined in advance. "Independently the agencies push out tentacles, brandishing separate RFP's. Firms on the other side of the chasm send out tentacles in response and, as on the Sistine ceiling, a spark leaps across" (Cronbach & Associates, 1980; quote from prepublished manuscript, p. 463). The contracting process itself shapes the evaluation as Baker (1975) discusses:

> Many applied research administrators push for such a detailed specification of the problem and research design that the only important question left for the contractor is how much it will cost to carry out the agency's plan. The agency, knowing what it wants done and how it wants it done, is looking for a skilled staff to carry out its needs, not somebody else's desires.
>
> The agency's desire to maximize control over the research, to make sure its problems get addressed the way the agency thinks [they] should be addressed, is precisely the reason why it uses contracts rather than grants. The important feature of a contract is that it maximizes the agency's control. (p. 210)

> The RFP is very important in the research process. It fixes the outline and many of the details of the study's methodology as well as specifying the problem to be studied. The RFP will generally define the population to be studied, sample sizes, and whether the study will be experimental, post-hoc interviews, or pre and post-field observations. The RFP may even specify the instruments to be used and the type of statistical analysis to be employed. In general, the two areas

> where the RFP leaves greatest discretion to the proposer is in the instrument content (the specific items) and data analysis. Note again that the RFP is prepared by the agency. The people who ultimately do the work have no involvement in many of the basic decisions of the research process. (Baker, as cited in Cronbach & Associates, 1980, p. 324).

There is room for negotiation; however, this pattern encourages the adoption of research proposals and designs without sufficient scrutiny. The day to day operations of the research corporation described by Cronbach and Associates (1980, p. 328), where there are plenty of "mouths to feed," provides an insight into the research corporation's behavior in this regard.

> Life in the contracting firm is dominated by the scramble for contracts. At every turn new money must be won to keep a staff in place. However, only large and experienced organizations can successfully solicit and manage large evaluations. A stack of blue chips is required merely to enter the bidding. The competitor must have a sophisticated business office for preparing proposals and keeping track of expenses. A public-relations staff stands by, ready to protect the flanks of a politically sensitive study. Computer facilities have to be extensive and up-to-date. Professional managers are needed to keep activities on schedule. And behind the scenes the firm's Washington representative keeps in touch with those who will be commissioning evaluations. Abert (quoted in Biderman & Sharp, 1972, p. 49) commented cynically that good research directors are far less necessary to a firm's success than are intelligence agents able to pick up early word on bidding opportunities. But the firm does what it can to maintain a staff of professionals qualified to plan, collect, and interpret data.
>
> Some firms offer services of many kinds, in many program areas. Once well established, a diversified firm can take the ups and downs of fortune more easily than a specialized firm. But even the largest firm shivers during a budget freeze, and it goes into a spasm of readjustment when it wins an unusually large contract. A narrow specialty makes an organization highly sensitive to the funding priorities of agencies. Over and over the same tale is told. A firm waxes as federal interest in its specialty grows. It welds together a team with complementary skills. The team accumulates special knowledge of the social problem. Then support disappears, the team splits up, and a capable organization is lost. (Abt, 1979, p. 50).

Excessive protests regarding the study's design jeopardize the corporation's chances of winning a contract. The business orientation promotes compromises, which may contribute to the overall pattern of misused designs. In addition, researchers' are socialized according to the same canons of educational research as the federal bureaucrats—the educational research establishment.

The educational research establishment's orientation is an even more powerful influence contributing to the repeated misuse of this paradigm. This view is characterized by the experimental, quantitative approach to research. Campbell and Stanley (1963), and Riecken et al. (1974) are probably the most widely recognized proponents of this approach. Campbell and Stanley wrote in their seminal work, *Experimental and Quasi-Experimental Designs for Research*:

> This chapter is committed to the experiment: as the only means for settling disputes regarding educational practice, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a fadish discard of old wisdom in favor of inferior novelties. (p. 2)

The fundamental elements of the paradigm are treatment and control groups, such as those used in the demonstration study under discussion.

Traditional educational researchers dominate evaluation research corporations. They have been socialized in graduate training to accept this orthodox credo. Educational researchers employing alternative methods or perspectives are regarded as operating outside the mainstream of "acceptable" educational research. An overemphasis on the importance of the design has led to a situation in which the methodological tail wags the proverbial research dog. Researchers such as those in the demonstration study have

allowed specific tools to dictate the way research would be conducted, rather than identifying the research questions and then selecting the appropriate method required to respond to them. This was, however, partially a function of the federal dictates.

The author is aware of the recent modifications made by some of the leading proponents of the educational research establishment. For example, Campbell (1979) has written in "an extreme oscillation away from [his] earlier dogmatic disparagement of case studies" that

> We should recognize that participants and observers have been evaluating program innovations for centuries without the benefit of quantification or scientific method. This is the common-sense knowing which our scientific evidence should build upon and go beyond not replace. But it is usually neglected in quantitative evaluations, unless a few supporting anecdotes haphazardly collected are included. Under the epistemology I advocate, one should attempt to systematically tap all the qualitative common sense program critiques and evaluations that have been generated among the program staff, program clients and their families, and community observers. While quantitative procedures such as questionnaires and rating scales will often be introduced at this stage for reasons of convenience in collecting and summarizing, nonquantitative methods of collection and compiling should also be considered, such as hierarchically organized discussion groups. Where such evaluations are contrary to the quantitative results, the quantitative results should be regarded as suspect until the reasons for the discrepancy are well understood. Neither is infallible, of course. But for many of us, what needs to be emphasized is that the quantitative results may be as mistaken as the qualitative. (pp. 52–53)

There is, however, a time lag between the deeply engrained socialization patterns of the past and the acceptance of new ideas and views emanating from the center of the educational research establishment. The world of contract research is somewhat removed from and often antagonistic to the halls of academia—the center of the educational research establishment—and requires additional time for the diffusion of new ideas.

Moreover, the same Campbell and Stanley "hard line" approach described earlier is highly visible in governmental circles today, as evidenced by a major document produced by Boruch and Cordray (1980): *An Appraisal of Educational Program Evaluations: Federal, State, and Local Agencies.* Boruch and Cordray, in their Executive Summary for Congress, recommended that "the higher quality evaluation designs, especially randomized experiments, be authorized explicitly in law for testing new programs, new variations on existing programs, and new program components" (first page of publication). This position is repeated throughout the document, for example, in the Executive Summary for the Department of Education, in a discussion on randomized field experiments, and so on. Their rationale for this recommendation parrallels that proposed by Tukey, Gilbert et al., Campbell and Stanley, among others.

> The main justification is that high quality designs lead to far less debatable estimates of programs on children than low quality designs. They are more difficult to execute, and they are more feasible for pilot testing new programs, program variations, and program components, than for estimating the effects of ongoing programs. Explicit authorization would make the importance of good designs plain, and would provide more clear opportunity for competent SEAs and LEAs to exploit them (Boruch & Cordray, 1980, p. 6).

This kind of justification is not valid when the application is either unethical or when the "no-cause" baseline is not established as in the national evaluation study under discussion. Furthermore, mandating stereotypic evaluation designs or paradigms is at best "off target." The focus on internal validity is misleading; external validity is the crux of the argument. "Internal validity . . . is not of salient importance in an evaluation. What counts in evaluation is external validity, the plausibility of conclusions about one or another *UTO that is significant to the PSC" (Cronbach & Associates, 1980, p. 314).

An analysis of the federal bureaucracy,

the research corporation, and the educational research establishment perspectives demonstrates how these parties can produce unintentionally undesirable effects (or "treatments") on program operations; effects that cannot be separated from the evaluation of the program and related "outcomes." The application of the holistic perspective demands that our attention be drawn to the policy context of the program and program evaluation. This perspective, like Dali's painting of Dali's painting ad infinitum, focuses on the importance of stepping back from the canvas to gain a more complete perspective of the portrait.

The generation of this demonstration project's research design and its acceptance are appropriate patterns of behavior given the real world constraints and perspectives discussed above; however, the behaviors dictated by these orientations and conditions often inhibit, rather than foster, the appropriate use of research paradigms.

Conclusion

Governmental agencies have traditionally equated the most credible research with the employment of the experimental design, regardless of the nature of the task. Ethnographic evaluations are novel innovations that are regarded at present as secondary to traditional, quantitative approaches. The traditional approach is adopted to make the strongest case before Congress—on whom they depend for future funds. This design is selected in accordance with the traditional canons of the educational research establishment. The federal climate of inflexible deadlines, interagency rivalry, and scarce resources forces bureaucrats to find the most convincing design for their audience, for their own political and economic survival. The federal bureaucrats then prepare the requests for proposal for research corporations, who in turn respond to federal interests. Therefore, we come full circle. The researchers are responsible for implementing the design as well as responding to RFPs, which explicitly or implicitly require the employment of a specific research design, regardless of the task at hand.

There has been considerable disillusionment with quantitative methods particularly with the experimental approach. Campbell and Stanley (1963), note that:

> Good and Scates (1954, pp. 716–721) have documented a wave of pessimism, dating back to perhaps 1935, and have cited even that staunch advocate of experimentation, Monroe (1938), as saying "the direct contributions from controlled experimentation have been disappointing." (p. 2)

This disillusionment has also been extended to the use of the experimental design in educational evaluation (Cronbach & Associates, 1980; Scriven, 1978; R. S. Weiss & Rein, 1969; C. H. Weiss, 1974, among others). In fact, governmental agencies, most notably the National Institute of Education, have funded several qualitative evaluation studies over the past 5 years in response to the problems arising from the application of experimental design to natural social settings. These awards may represent a shift in allegiances regarding paradigms. This discussion itself may exist within the content of a scientific revolution of paradigms in educational evaluation: qualitative versus quantitative. Kuhn (1962) explained that the acceptance of a new paradigm depends on prior crisis, faith, and many arguments.

> The man who embraces a new paradigm at an early stage must often do so in defiance of the evidence provided by problem-solving. He must, that is, have faith that the new paradigm will succeed with the many large problems that confront it, knowing only that the older paradigm has failed with a few. A decision of that kind can only be made on faith.
>
> That is one of the reasons why prior crisis proves so important. Scientists who have not experienced it will seldom renounce the hard evidence of problem-solving to follow what may easily prove and will be widely regarded as a will-o'-the-wisp. But crisis alone is not enough. There must also be a basis, though it need be neither rational nor ultimately correct, for faith in the particular candidate chosen . . .
>
> This is not to suggest that new paradigms triumph ultimately through some mystical aesthetic. On the contrary, very few men desert a tradition for these reasons alone. Often those who do turn out to have been misled. But if a paradigm is ever to triumph it must gain some first

> supporters, men who will develop it to the point where hardheaded arguments can be produced and multiplied. And even those arguments, when they come, are not individually decisive. Because scientists are reasonable men, one or another argument will ultimately persuade many of them. But there is no single argument that can or should persuade them all. Rather than a single group conversion, what occurs is an increasing shift in the distribution of professional allegiances. (p. 158)

Kuhn's conversion experience does not occur overnight. It is not unusual to observe "lifelong resistance particularly from those whose productive careers have committed them to an older tradition of normal science" (p. 151). In addition, because each paradigm has elements of the other, the new paradigm will probably be a Hegelian synthesis of the two contrasting paradigms, rather than a dominance of one over the other.

Currently, however, the dominant research mode of sponsors, managing agencies, and the educational research establishment follows the traditional quantitative orientation. The quintessential paradigm of this orientation is the experimental design.

Both quantitative and qualitative methods are presently required to answer different kinds of questions—and aid each other in the same questions—in evaluation research. (See Cronbach & Associates, 1980; Campbell, 1974; R. S. Weiss & Rein, 1969, for discussions of the use of qualitative data and interpretation in evaluation). The misapplication of the experimental design on a national level does little to stem the tide of disillusionment with the quantitative mode. The continued misuse of this paradigm will only render it impotent, in much the same manner that the misuse of the qualitative paradigm will render it a fad in educational evaluation. I am inclined to agree with a colleague of mine, that "this disillusionment (with the quantitative method) is misplaced and the product of poor understanding of what different methods do and do not try to do". Research paradigms, however, require sensitivity and proper application to clients whether quantitative or qualitative. The misapplication of either design requires attention and examination.

The value of the experimental design can be compared to the value of technology—it is neither good nor bad, useful nor useless per se; only specific applications are good or bad, useful or useless. The repeated misuse of the experimental design is a function of several mutually reinforcing perspectives and real world constraints. Paradigms, in theory, do not logically determine the choice of research methods, as Reichardt and Cook have demonstrated (1979, pp. 11-32). In practice, however, paradigms do lend themselves to the use of one research method more readily than another. The author supports the increased use of qualitative methods, specifically ethnographic techniques, in social policy research.[3] These techniques serve to respond more appropriately to certain evaluation concerns (e.g., process evaluation). Ethnographic "evaluations assume that human institutions are multi-dimensional and that social interventions have multiple facets and multiple relationships with multiple results" (Britan, 1980, p. 5). In addition, the use of qualitative methods

[3] Quantitative and qualitative mixture of expertise in federal and state should continue. Ethnographic research can prevent many of the problems discussed. Preliminary ethnographic research can be conducted at the beginning of the evaluation to establish a valid baseline (regarding the nature of the program) and generate numerous hypotheses related to both program participant and legislator concerns. In addition, this information would be useful in the development and choice of appropriate instruments during the "summative" component of the evaluation. This information would also enable evaluators to determine the importance or value of specific program outcomes. For example, the ethnographic research may establish that the basic aim of the program is to produce attitudinal changes of a specific nature. Standardized reading and math tests would be useful if it is an educational program; however, it would be understood that these were secondary or tertiary aims of the program. Similarly, appropriate instruments required to measure a specific kind of attitudinal change could be appropriately selected for the "summative" evaluation. (There are various reliable indices of attitudinal change that do not require the use of standarized instruments, e.g., attendance, observed activities, informal interviews, semistructured interviews, systematic alterations in apparel, language, and behavior of participants and various external manifestations). Finally, community concerns regarding evaluation techniques can be discerned and addressed before various pressures arise (e.g., boycotting a program).

serves to interrupt the chain of reinforcing perspectives that often blind practitioners to the task at hand. This does not suggest, however, that research strategies deriving from the qualitative world view are "superior to experimental design as a methodology for evaluating broad-aim programs" as is argued by others (e.g., R. S. Weiss & Rein, 1972, p 243), nor that extreme opposition to the quantitative approach (Hamilton et al., 1978) is required or useful.

The appropriate use of both qualitative and quantitative design are required to respond properly to social policy research, as discussed above. Campbell's call for a clearer understanding of the relationship between qualitative and quantitative ways of knowing will contribute to an understanding of the larger problems facing educational research. Discussions of this nature reveal this chain of reinforcing real-world constraints and views, and thus allow us to break away from this chain and view the task at hand more clearly.

The problem on one level has been the simple misapplication of the treatment-control design. On another level, the problem is the power of reinforcing constraints and world views that generate maladaptive behavior.

Ibsen's (1958) Stockmann discovered, after being rebuked for his attempts to publish his findings regarding the contaminated baths, that "all the sources of [their] moral life [were] poisoned and that the whole fabric of [their] community [was] founded on the pestiferous soil of falsehood" (p. 653). The discovery presented in this discussion is that our research community in its efforts to produce the best research results is methodologically blinded by the very world view that generates one of the most cherished designs in social science.

This discussion calls for a reexamination of paradigms, research practices and policies, as well as the underlying real-world constraints and views that generate them. The danger of narcissistic reflection exists in the realm of social science. The unexamined self, however, represents a greater danger, threatening the whole fabric of our community—like Ibsen's baths.

Recommendations

The recommendations presented below are directly and indirectly related to the federal study under discussion. In addition, they both support and criticize the recommendations posited by Cronbach and Associates (1980).

(1) Abandon treatment-control designs in projects with disaffected, "disadvantaged," or disenfranchised populations. The application of this design has unintentionally deleterious effects.

a. Depriving individuals of an opportunity to enter the mainstream of society may cost more than the knowledge gained from a controlled trial.

b. There are serious methodological flaws in an evaluation that applies this design to these populations and programs, for example, reactive arrangements, Hawthorne effect, John Henry effect.

c. This design sheds an unfavorable light on the sincerity of the project to many community members, LEA officials, and the like. This is associated with previously existing perceptions of demonstration projects as "rip-offs" that "come and go . . . getting our hopes up and then letting us down." In fact, these associations and perceptions may contribute to recruitment difficulties experienced currently by demonstration projects throughout the country. Recruitment may represent a test of how communities are reacting to these types of demonstration projects (with their treatment and control constraints) rather than a test of how the community likes a specific program per se.

(2) "A particular control is warranted if it can be installed at reasonable cost [including the social costs discussed above], and if, in the absence of that control, a positive effect could be persuasively explained away" (Cronbach & Associates, 1980, p. 552).

(3) Multiple indicators of outcomes should be used to assess program implementation and changes in program participant behavior.

(4) External validity—the ability to generalize from the data—is the key to a use-

ful evaluation. Technical quality particularly in the form of an overemphasis on internal validity does not increase utilization. (See Caplan, 1977, pp. 187–188; Patton, 1977, p. 151).

(5) Evaluation research can become more useful if guided by a knowledge of the dynamics of policymaking.

(6) Conclusions about a program require knowledge of the context of the program. "Stand alone" studies rarely draw on the numerous sources of information available in an evaluation. In addition, they draw their conclusions about a program without regard for the context of the program and the evaluation itself. "It is better for an evaluation inquiry to launch a small fleet of studies than to put all its resources into a single approach" (Cronbach & Associates, 1980, p. 7).

(7) Replication is a biological not an anthropological or sociological concept. Programs adapt to their environment. The process of adaptation should be the focus of inquiry (see Fetterman, 1981).

(8) Evaluators should use both quantitative and qualitative research methodologies in evaluation. In addition, evaluators should declare their allegiance to methodological persuasions based on formal training, experience, and disposition. Ideally, allegiance to a methodology is superfluous. Given the political and ideological realities of evaluation, however, it is unrealistic to expect qualitative methodologies to be properly represented. Moreover, the fashionability of the methodology has attracted many sympathetic but untrained supporters. Individuals properly trained and suited to conduct ethnographic research in a contract research setting, however, are required if ethnography is to continue to contribute to the field of evaluation. (See Wolcott, 1975, as a guide to the selection of ethnographers and the conduct of ethnographic research in school settings). Formal declarations based on training and experience will contribute to the proper representation and use of the methodology in evaluation.

(9) Evaluation should be primarily used to understand program dynamics and outcomes. A legitimate secondary or latent function of evaluation, however, is accountability. The presence of an evaluation itself serves to ensure that basic program implementation requirements are met. Moreover, accountability is a political reality. In addition, there is a moral obligation involved in holding programs and evaluations accountable in the use of taxpayer monies. Calls for accountability, however, can be abused. Evaluations should not be used in order to "praise or blame" programs.

(10) "Goal based" evaluations have been misleading. Goals are often part of the political rhetoric, or they are misunderstood, or poorly determined and therefore misrepresent the program. All social programs have broad aims. Narrowly focusing on whether a program has attained its prespecified goals is rarely productive.[4] Using Evaluability Assessment (see Wholey, 1979) procedures, formative evaluation procedures, or ethnographic techniques (see Fetterman, 1980), however, can contribute to a more accurate understanding of the various aims of a program, as well as manifest and latent functions of the program.

(11) Increased sensitivity to the unintentional effects of research corporation, educational research establishment, and policymaker real-world constraints and views in program operations could mitigate future, unintentionally deleterious treatments on program operation and evaluation.

a. Greater sensitivity and effort is required of evaluators and policy makers to discern whether a specific design is appropriate for an evaluation in the request for proposals. An atmosphere for discussing alterations must exist between monitor and contractor if researchers and sponsors are expected to admit that they don't know everything.

(12) Evaluators have an ethical obligation to share their information with all parties concerned. Prespecified periods should be established for disclosures at the beginning of the contract and amenable to significant developments in the evaluation process. Releasing findings in a completely "piecemeal" fashion is more often than not

[4] Scriven's (1978) dichotomy of formative and summative evaluation are useful pedagogical tools for understanding evaluation; however, in practice all evaluation is of a formative nature.

disruptive, misleading, misused, and rarely equitable in its parcels.

(13) "Communication overload is a common fault; many an evaluation is reported with self-defeating thoroughness" (Cronbach & Associates, 1980, p. 6). It is true that "When an avalanche of words and tables descends, everyone in its path dodges" (1980, p. 186).[5] This recommendation, however, should not dampen efforts to complete a thorough, scholarly, investigation. Long, detailed findings can be presented in a technical report, while a summary of salient findings or a nontechnical report can be produced for policymakers.

(14) An artificial deadline set at the beginning of a study should not dominate the study when there is no specific time at which the information is required to make policy decisions. Realistically, however, evaluators must respect a sponsoring agencies timely constraints when a policy decision does rely on the timely input of information—if they are to be useful to policymakers and if they are to continue working.

(15) Multiple sponsorship or interagency agreements can bring different perspectives to bear on an evaluation; however, interagency rivalry and poor lines of communication result more often than not from such an endeavor, severely affecting program operations. It is often like two big game animals fighting to display who is king of the jungle and when "the dust settles it is the earth that loses" (Fetterman, 1981).

(16) Peer review is a common policy in evaluation research—formally and informally. This policy should continue; however, more rigorous criticisms are required.

References

ABT, C. C. Government constraints on evaluation quality. In L. E. Datta & R. Perloff (Eds.), *Improving educations*. Beverly Hills, Calif: Sage, 1979.

ANDERSON, R. B., ET AL. Pardon us, but what was the question again? A response to the critique of the follow-through evaluation. *Harvard Educational Review*, 1978, *48*, 161-170.

ACLAND, H. Are randomized experiments the Cadillacs of design? *Policy Analysis*, 1979, *5*, 223-241.

BAILEY, F. *Tribe, caste, and nation*. Manchester, England: Manchester University Press, 1960.

BAKER, K. A new grantsmanship. *American Sociologist*, 1975, *10*, 206-219.

BARTH, F. *The role of the entrepreneur in social change*. Northern Norway, Bergen: Scandinavian University, 1963.

BARTH, F. *Models of social organization*. Occasional Paper No. 23. Royal Anthropological Institute of Great Britain and Ireland, 1966.

BARTH, F. On the study of social change, *American Anthropologist*, 1967, *69*(6), 661-669.

BIDERMAN, A. D., & SHARP, L. M. *The competitive evaluation research industry*. Washington, D.C.: Bureau of Social Science Research, 1972.

BORUCH, R. F. On common contentions about randomized field experiments. Originally in Boruch, R. F., & Riecken, H. W. *Experimental testing of public policy: The Proceedings of the 1974 Social Science Research Council Conference on Social Experiments*. Boulder, Colo.: Westview Press, 1975. (Cited from *Evaluation Studies Review Annual*, 1976, *1*, 158-193.)

BORUCH, R. F., & CORDRAY, D. S. *An appraisal of educational program evaluations: Federal, state, and local agencies*. Evanston, Ill: Northwestern University, 1980.

BRITAN, G. *Evaluation as science or politics: The role of ethnography in policy change*. Paper presented at the annual meeting of the American Anthropological Association, Washington, D.C., December 1980.

BRONFENBRENNER, U. The experimental ecology of education. *Teachers College Record*, 1976, *78*, 157-204.

BYAR, D. P., SIMON, R. M., FRIEDEWALD, W. T., SCHLESSELMAN, J. J., DEMETS, D. L., ELLENBERG, J. N., GAIL, M. H., WARE, J. H. Randomized clinical trials: Perspectives on some recent ideas. *New England Journal of Medicine*, 1976, *74*, 295.

CAMPBELL, D. T. Qualitative knowing in action research. Occasional paper. Stanford Evaluation Consortium, Stanford University, 1974.

CAMPBELL, D. T. Degrees of freedom and the case study. In T. D. Cook & C. S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research*. Beverly Hills, Calif.: Sage, 1979.

CAMPBELL, D. T., & STANLEY, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1963.

CAPLAN, N. A minimal set of conditions necessary for the utilization of social science knowledge in policy formulation at the national level. In C. H. Weiss (Ed.), *Using social research in public policy making*. Lexington, Mass.: Lexington Books, 1977.

COMPTROLLER GENERAL. *Assessing social program impact evaluations: A checklist approach*. Washington, D. C.: General Accounting Office, 1978.

COOK, T. D., & CAMPBELL, D. T. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.

[5] Cronbach and Associates present an interesting example of "self-defeating thoroughness" in evaluation. "As an extreme example, one recent report on nonformal education in Latin America ran to 900 pages (single spaced and in English!); perhaps that document found no audience at all" (1980, p. 186).

COWARD, R. The involvement of anthropologists in contract evaluations: The federal perspective. *Anthropology and Education Quarterly*, 1976, *7*, 12–16.

CRONBACH, L. J. Five decades of public controversy over mental testing. *American Psychologist*, 1975, *30*, 1–13.

CRONBACH, L. J., AMBRON, S. R., DORNBUSCH, S. M., HESS, R. D., HORNIK, R. C., PHILLIPS, D. C., WALKER, D. F., & WEINER, S. S. *Toward reform of program evaluation: Aims, methods, and institutional arrangements*. San Francisco: Jossey-Bass, 1980.

ELISBURG, D. A Congressional view of program evaluation. In E. Chelimsky (Ed.), *A Symposium on the use of evaluations by federal agencies* (Vol. 1). McLean, Va.: Mitre Corporation, 1977.

ELMORE, R. F. Lessons from follow-through. *Policy Analysis*, 1975, *1*, 459–484.

ERASMUS, C. *Man takes control*. Indianapolis, Ind.: Bobbs-Merrill, 1961.

ETZIONI, A. Policy research. *American Sociologist*, 1971, *6*, 8–12, (Supplemental Issue).

FETTERMAN, D. M. Ethnographic techniques in educational evaluation: An illustration. Special topic edition of the *Journal of Thought*, Anthropology of Education: Methods and Applications, 1980, *15*(3), 31–48.

FETTERMAN, D. M. Blaming the victim: The problem of evaluation design, federal involvement, and reinforcing world views in education. *Human Organization*, 1981, *40*, 67–77.

GILBERT, J. P., MCPEEK, B., & MOSTELLER, F. Statistics and ethics in surgery and anesthesia. Originally in *Science*, 1977, *198*, 679–684. (Cited from *Evaluation Studies Review Annual*, 1978, *3*, 333–338).

GLASS, G. V Policy for the unpredictable (uncertainty research and policy). *Educational Researcher*, 1979, *8*(9), 12–14.

GLENNAN, T. K. Evaluating federal manpower programs: Notes and observations. In P. H. Rossi & W. Williams (Eds.), *Evaluating social programs: Theory, practice, and politics*. New York: Seminar Press, 1972.

GOOD, C. V., & SCATES, D. E. *Methods of research*. New York: Appleton-Century-Crofts, 1954.

HANEY, W. *A technical history of the National Follow-Through Evaluation*. Cambridge, Mass.: Huron Institute, 1977.

HAMILTON, D. (Eds.). *Beyond the numbers game*. Berkeley, Calif.: McCutchan, 1978.

HOLCOMB, H. Tell Congress results of research. *Education Daily*, 1974, *7*(4), 313.

IBSEN, H. An enemy of the people. In R. B. Inglis & W. K. Steward (Eds.), *Adventures in world literature*. New York: Harcourt, Brace, and World, 1958.

KUHN, T. S. *The structure of scientific revolutions*. Chicago: The University of Chicago Press, 1962.

KUNKEL, J. *Society and economic growth*. New York: Oxford University Press, 1970.

LINDBLOM, C. E., & COHEN, D. K. *Usable knowledge*. New Haven, Conn.: Yale University Press, 1979.

MARCH, J. G., & OLSEN, J. P. *Ambiguity and choice in organizations*. Bergen, Norway: Harald Lyche, 1976.

MATARAZZO, J. Psychotherapeutic processes. *Annual Review of Psychology*, 1965, *16*, 181–224.

MCCLELLAND, D., & WINTER, D. *Motivating economic achievement*. New York: The Free Press, 1969.

MONROE, W. S. General methods: Classroom experimentation. In G. M. Whipple (Ed.), *Yearbook of the National Society of Studies in Education* (Part II. Vol. 37), 1938.

MULHAUSER, F. Ethnography and policymaking: The case of education. *Human Organization*, 1975, *34*, 311.

PATTON, M. L. ET AL. In search of impact: An analysis of the utilization of federal health evaluation research. In C. H. Weiss (Ed.), *Using social research in public policy making*. Lexington, Mass.: Lexington Books, 1977.

PETO, R., PIKE, M. D., ARMITAGE, P., BRESLOW, N. E., COX, D. R., HOWARD, S. V., MANTEL, N., MCPHERSON, K., PETO, J., & SMITH, P. G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I, Introduction and design, *British Journal of Cancer*, 1977, *34*, 585. (For part II see ibid., 1977, *35*, 1).

PILLEMER, D. E., & LIGHT, R. J. Using the results of randomized experiments to construct social programs: Three caveats. In L. Sechrest et al. (Eds.), *Evaluation Studies Review Annual*, 1979, *4*, 717–726.

REICHARDT, C. S., & COOK, T. D. Beyond qualitative versus quantitative methods. In T. D. Cook & C. S. Reichardt (Eds), *Qualitative and quantitative methods in educational research*. Beverly Hills, Calif.: Sage Publications, 1979.

REICKEN, W. R., BORUCH, R. F., CAMPBELL, D. F., CAPLAN, N., GLENAN, F. K., JR., PRATT, J. W., REES, A., & WILLIAMS, W. *Social experimentation: A method for planning and evaluating social intervention*. New York: Academic Press, 1974.

ROETHLISBERGER, F. J., & DICKSON, N. J. *Management and the worker*. Cambridge: Harvard University Press, 1939.

SARETSKY, G. The OEO P.C. experiment and the John Henry effect. *Phi Delta Kappan*, 1972, *53*, 579–581.

SCRIVEN, M. Two main approaches to evaluation. In R. M. Bossone (Ed.), *Proceedings, Second National Conference on Testing*. New York: Center for Advanced Study in Education, City University of New York, 1978.

TALLMADGE, G. K. Avoiding problems in evaluation. *Journal of Career Education*, 1979, *5*(4), 300–308.

TUKEY, J. W. Some thoughts on clinical trials, especially problems of multiplicity. Originally in *Science*, 1977, *198*, 679–684. (Cited from *Evaluation Studies Review Annual*, 1978, *3*, 327–332).

VON NEUMAN, J., & MORGANSTERN, O. *The theory of games and economic behavior* (3rd ed.). Princeton, N.J.: Princeton University Press, 1953.

WEISS, C. H. Alternative models of program evaluation. *Social Work*, 1974, *19*, 675–681.

WEISS, C. H. (Ed.). *Using social research in public policy making*. Lexington, Mass.: Lexington Books, 1977.

WEISS, R. S., & REIN, M. The evaluation of broad-aim programs: A cautionary case and a moral. *Annals of the American Academy of Political and Social Science*, 1969, *385*, 133–142.

WEISS, R. S., & REIN, M. The evaluation of broad-aim programs: Difficulties in experimental design and an alternative. In C. H. Weiss (Ed.), *Evalu-*

ation action programs: Readings in social action and education. Boston: Allyn & Bacon, 1972.

WHOLEY, J. *Evaluation promise and performance.* Washington, D.C.: The Urban Institute, 1979.

WOLCOTT, H. Criteria for an ethnographic approach to research in schools. *Human Organization,* 1975, 34(2), 111-127.

Author

DAVID M. FETTERMAN, Senior Associate, RMC Research Corp., 2570 W. El Camino Real, Mountain View, CA 94040; Project Director of California Arts Council—Evaluation; lecturer, Anthropology Dept., Stanford University, Stanford, CA 94305. *Specializations:* Educational and Medical Anthropology and Educational Policy Analysis.