EVALUATING FEDERAL EDUCATION PROGRAMS

Eva L. Baker, Editor

November 1980

CSE Report No. 153

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

PREFACE

This volume presents a series of papers developed by the authors as part of their deliberations as members on the National Research Council's Committee of Program Evaluation in Education. As such, the papers represent only the work of the individual authors and in no way have been reviewed or endorsed by the Committee, the National Research Council, the National Academy of Science, and while we're disclaiming, the UCLA Center for the Study of Evaluation, nor the National Institute of Education.

Yet while officially representing only themselves, the authors by topic and style provide a compact panoply of present evaluation thinking. Some authors hold the view, "let the flower bloom" and address their topic with dispassionate appraisal of alternatives. No less skillful writers on the other hand, eschew such even handedness and inform the reader of the characteristics of high quality studies.

Some papers are didactic, and attempt to take care to explain with little ambiguity or abstraction what the critical issues are. Others present an enticing range of examples and support for favorite theses and summon up work horizontally from other science fields and historically from earlier epochs. Lacing the writing so richly heightens the reader's perception of the breadth and credibility of the author. These various virtues can be found within these efforts.

Certainly, if never plainly detailed, the conflict between preferences in evaluation methodology comes through in these papers. The problem is not the oft contrasted benefits of quantitative versus qualitative information. Rossi and Berk correctly point out that various methods of data generation can be used complementarily in both experimentally controlled and naturally varying designs. Left for future debate, however, is the avowal of some of the authors that experimental manipulation represents the evaluation design of choice. Rather casually dismissed were concerns of political solvency, of program diffusion (or contamination), conservatism and delay inherent in developmental testing as these concerns might specially impact program evaluation choices in education. For myself, I would have enjoyed an extended discussion of the measurement problems in any of the designs available to us.

But my intention is to praise the diversity that these views communicate. One particular pleasure in reading these papers derived from the evaluation in contexts used by the writers. Rather than arguing as solely evaluators on the issue of program evaluation, more than one writer found it necessary to explain the role of evaluation in the larger theater of educational development. Redirecting evaluators' views from exclusively research roots and extending their colloquium to include both research and development should certainly improve both the quality, academic virtue, and practicality of consequent evaluations.

Eva L. Baker
November, 1981

TABLE OF CONTENTS

CHAPTER 1

AN OVERVIEW OF EVALUATION STRATEGIES AND PROCEDURES

Peter H. Rossi
University of Massachusetts

Richard A. Berk
University of California, Santa Barbara

Introduction

The purpose of this paper is to provide a detailed introduction to the variety of purposes for which evaluation research may be used and to the range of methods that are currently employed in the practice of that field. Specific examples are given wherever appropriate to provide concrete illustrations of both the goals of evaluation researches and the methods used.

While the coverage of this paper is intended to be comprehensive in the sense of describing major uses of evaluation research, it cannot even pretend to be encyclopedic. The reader interested in pursuing any of the topics discussed in this paper is provided with references to more detailed discussions. In addition, there are several general references that survey the field of evaluation in a more detailed fashion (Suchman, 1967; Weiss, 1972, Cronbach, 1980, Rossi, Freeman & Wright, 1979; Guttentag & Struening, 1976).

Policy Issues and Evaluation Research

Virtually all evaluation research begins with one or more policy questions in search of relevant answers. Evaluation research may be conducted to answer questions that arise during the formulation of policy, in the design of programs, in the improvement of programs, and in testing the efficiency and effectiveness of programs that are in place or being considered. Specific policy questions may be concerned with how widespread a social problem may be, whether any program can be enacted that will ameliorate a problem, whether programs are effective, whether a program is producing enough benefits to justify its cost, and so on.

Given the diversity of policy questions to be answered, it should not be surprising that there is no single "best way" to proceed and that evaluation research must draw on a variety of perspectives and on a pool of varied procedures. Thus, approaches that might be useful for determining what activities were actually undertaken under some educational program, for instance, might not be appropriate when the time comes to determine whether the program was worth the money spent. Similarly, techniques that may be effective in documenting how a program is functioning on a day-to-day basis may prove inadequate for the task of assessing the program's ultimate impact. In other words, the choice among evaluation methods derives initially from the particular question posed; appropriate evaluation techniques must be linked explicitly to each of the policy questions posed. While this point may seem simple enough, it has been far too often overlooked, often resulting in force-fits between an evaluator's preferred method and particular questions at hand. Another result is an evaluation research literature padded with empty, sectarian debates between warring camps of "true believers". For example, there has been a long and somewhat tedious controversy about whether assessments of the impact of social programs are best undertaken with research designs in which subjects are randomly assigned to experimental and control groups or through theoretically derived causal models of how the program works. In fact, the two approaches are complementary and can be effectively wedded (e.g., Rossi, Berk, & Lenihan, 1980).

To obtain a better understanding of the fit between evaluation questions and the requisite evaluation procedures, it is useful to distinguish between two broad evaluation contexts, as follows:

> Policy and Program Formation Contexts: Contexts in which policy questions are being raised about the nature and amount of social problems, whether appropriate policy actions can be taken, and whether programs that may be proposed are appropriate and effective.
>
> Existing Policy and Existing Program Contexts: Contexts in which the issues are whether appropriate policies are being pursued and whether existing programs are achieving their intended effects.

While these two broad contexts may be regarded as stages in a progression from the recognition of a policy need to the installation and testing of programs designed to meet those policy needs, it is often the case that the unfolding of a program in actuality may bypass some evaluation activities. For example, Head Start and the Job Corps were started up with minimum amounts of program testing beforehand: indeed, the issue of whether Head Start was or was not effective did not surface until some years after the program had been in place, and the Job Corps has just recently (a decade and a half after enactment) been evaluated in a sophisticated way (Mathematica, 1980). Similarly, many programs apparently never get beyond the testing stage, either by being shown to be ineffective or troublesome (e.g., contract learning, Gramlich & Koshel, 1975) or because the policy issues to which they were addressed shifted in the meantime (e.g., as in the case of negative income tax proposals, Rossi & Lyall, 1974).

Unfortunately, a statement that evaluation techniques must respond to the questions that are posed at different stages of a program's life history, only takes us part of the way. At the very least, it is necessary to specify criteria that may be used to select appropriate evaluation procedures, given one or more particular policy questions. For example, randomized experiments are an extremely powerful method for answering some of the questions posed as part of program design issues, but may be largely irrelevant to or ineffective for answering questions associated with program implementation issues. Yet, such terms as "powerful", "relevant", and "ineffective" are hardly precise and the following four criteria are far more instructive.

First, one must consider whether the measurement procedures that are being proposed are likely to capture accurately what they are supposed to measure. Sometimes such concerns with measurement quality are considered under the rubric of "construct validity" (Cook & Campbell, 1979), and are germane to <u>all</u> empirical work regardless of the question being asked. For example, while it is apparent that an examination of the impact of <u>Sesame Street</u> on children's reading activity must rest on measures that properly reflect what educators mean by an ability to

read, the same concerns are just as relevant in ethnographic accounts of how parents "encourage" their children to watch Sesame Street. One must, presumably, have a clear idea of kinds of inducements parents might provide and field work procedures that systematically and accurately record the use of these inducements. At the very least, field workers would have to be instructed about how to recognize an "inducement" as distinct from other sorts of interaction occurring between parents and children.

It is important to stress that questions about measurement quality apply not only to program outcomes such as "learning", but also to measures of the program (intervention) itself and to other factors that may be at work (e.g., a child's motivation to learn). For example, the crux of the ongoing debate about the Hawthorne Experiments undertaken over 50 years ago involves a judgment of whether the "real" treatment was a physical alteration in the worker's environment or changes in worker-employer relations affecting employee motivation (Franke & Kaul, 1978; Franke, 1979; Wardwell, 1979).

Finally, while space limitations preclude a thorough discussion of measurement issues in evaluation research, two generic kinds of measurement errors should be distinguished. On one hand, measurement may be subject to bias that reflects a systematic disparity between indicator(s) and an underlying true attribute that is being gauged. Perhaps the most visible example is found in the enormous literature on whether standardized IQ tests really tap "general intelligence" in a culture free manner. (For a recent review see Cronbach, 1975.) On the other hand, measures may be flawed because of random error or "noise". Whether approached as an "errors in variables" problem through the econometric literature (e.g., Kmenta, 1971: 309-322) or as the "underadjustment" problem in the evaluation literature (e.g., Campbell & Erlebacher, 1970), random error can lead to decidedly non-random distortions in evaluation results. (For a recent discussion see Barnow, Cain & Goldberger, 1980.) The role of random measurement error is sometimes addressed through the concept of "reliability".

Secondly, many evaluation questions concern causal relations, as, for example, whether or not a specific proposed program of bilingual education will improve the English language reading achievement of program participants, i.e., whether exposure to a program will "cause" changes in reading achievement. Whenever a causal relationship is proposed, alternative explanations must be addressed and presumably discarded. If such alternatives are not considered, one may be led to make "spurious" causal inferences; the causal relationship being proposed may not in fact exist. Sometimes this concern with spurious causation is addressed under the heading of "internal validity" (Cook & Campbell, 1979) and, as in the case of construct validity, is relevant regardless of stage in a program's life history (assuming causal relationships are at issue). For example, no one would dispute that a causal inference that Head Start, for example, improves the school performance of young children, must also consider the alternative explanation that children participating in a Head Start program were "simply" brighter than their peers to begin with. Note that the same issues surface in ethnographic accounts of how programs like Head Start function. A series of documented observations suggesting that Head Start provides a more supportive atmosphere in which students may learn academic skills, requires that alternative explanations be considered. Thus, it may be that the content of Head Start programs are less important than the kinds of instructors who volunteer (or are recruited) to teach in such programs.

The consideration of alternative causal explanations for the working of social programs is an extremely important research design consideration. For example, programs that deal with humans are all subject more or less to problems of self-selection; often enough persons who are most likely to be helped or who are already on the road to recovery are those most likely to participate in a program. Thus, vocational training offered to unemployed adults is likely to attract those who would be most likely to improve their employment situation in any event. Or sometimes program operators "cream the best" among target populations to participate in programs, thereby assuring that such

programs appear to be successful. Or, in other cases, events unconnected with the program produce improvements which appear to be the result of the program: an improvement in employment for adults, for instance, may make it more likely that young people will stay in and complete their high school training.

It cannot be overemphasized that parallel design issues necessarily surface in evaluations based on qualitative field work. While this point has a long and rich history in the social sciences that routinely collect and analyze qualitative data (e.g., Zelditch, 1962; Becker, 1958; Mensh & Henry, 1953), evaluation researchers have to date been somewhat slow to catch on. Too often "process research", for example, has become a license for research procedures that are little more than funded voyeurism. In short, there is more to field work than simply "hanging out".

Third, whatever the empirical conclusions resulting from evaluation research during any of the three program stages, it is necessary to consider how broadly one can generalize the findings in question: that is, are the findings relevant to other times, other subjects, similar programs and other program sites? Sometimes such concerns are raised under the rubric of "external validity" (Cook & Campbell, 1979), and again, the issues are germane in all program stages and regardless of evaluation method. Thus, even if a quantitative assessment of high school driver education programs indicates that they do not reduce the number of automobile accidents experienced by teenagers (Roberts, 1980), it does not mean that adult driver education programs would be ineffective. Similarly, an ethnographic account of why the driver education program did not work for teenagers, may or may not generalize to adult driver education programs.

Generalization issues ordinarily arise around several types of extensions of findings. For instance, are the findings applicable to other cities, agencies, or school systems, besides the ones in which they were found? Or are the results specific to the organizations in which the program was tested? Another issue that arises is whether a program's results would be applicable to students who are different in abilities or in socioeconomic background? For example, Sesame Street was found

to be effective with respect to preschool children from lower socio-economic families, but also more effective with children from middle class families (Cook, et al., 1975). Or, curricula that work well in junior colleges may not be appropriate for students in senior colleges. There is also the problem of generalizing over time. For example, Maynard and Murnane (1979) found that transfer payments provided by the Gary Income Maintenance Experiment apparently increased the reading scores of children from the experimental families. One possible explanation is that with income subsidies, parents (especially in single parent families) were able to work less and therefore spend more time with their children. Even if this is true, it raises the question of whether similar effects would be found presently when inflation is taking a much bigger bite out of the purchasing power of households. Finally, it is impossible to introduce precisely the same treatment(s) when studies are replicated or when programs move from the development to the demonstration stage. Hence, one is always faced with trying to generalize across treatments that can rarely be identical. In summary, external validity surfaces as a function of the subjects of an evaluation, the setting, the historical period and the treatment itself. Another way of phrasing this issue is to consider that programs vary in their "robustness"; that is, in their ability to produce the same results under varying circumstances, with different operators, and at different historical times. Clearly a "robust" program is highly desirable.

Finally, it is always important to consider that whatever one's empirical assessments, that the role of "chance" is properly taken into account. When formal, quantitative findings are considered, this is sometimes addressed under the heading of "statistical conclusion validity" (Cook & Campbell, 1979) and the problem is whether tests for "statistical significance" have been properly undertaken. For example, perhaps Head Start children appear to perform better in early grades, but at the same time, the observed differences in performance could easily result from chance factors having nothing to do with the program. Unless the role of these chance factors is formally assessed, it is impossible to determine if the apparent program effects are real

or illusory. Similar issues appear in ethnographic work as well as though formal assessments of the role of chance are difficult to undertake in such studies. Nevertheless, it is important to ask whether the reported findings rest on observed behavioral patterns that occured with sufficient frequency and stability to warrant the conclusions that they are not "simply" the result of chance. No self-respecting ethnographer would base an analysis of the role of parental inducements in impact of Sesame Street, for example, on a single parent-child interaction on a particular morning.

Three types of factors play a role in producing apparent (chance) effects that are not "real". The first reflects sampling error and occurs whenever one is trying to make statements about some population of interest from observations gathered on a subset of that population. For example, one might actually be studying a sample of students from the population attending a particular school, or a sample of teachers from the population of teachers in a particular school system, or even a sample of schools from a population of schools within a city, county, or state. Yet, while it is typically more economical to work with samples, the process of sampling necessarily introduces the prospect that any conclusions based on the sample may well differ from conclusions that might have been reached had the full population been studied instead. Indeed, one could well imagine obtaining different results from different subsets of the population.

While any subset that is selected from a larger population for study purposes may be called a sample, some such subsets may be worse than having no observations at all. The act of sampling must be accomplished according to rational selection procedures that guard against the introduction of selection bias. A class of such sampling procedures that yield unbiased samples are called "probability samples", in which every element in a population has a known chance of being selected (Sudman, 1976; Kish, 1965). Probability samples are difficult to execute and are often quite expensive, especially when dealing with populations that are difficult to locate in space. Yet there are such clear

advantages to such samples, as opposed to haphazard and potentially biased methods of selecting subjects, that probability samples are almost always to be preferred over less rational methods.

Fortunately, when samples are drawn with probability procedures, disparities between a sample and a given population can only result from the "luck of the draw," and with the proper use of statistical inference, the likely impact of these chance forces can be taken into account. Thus, one can place "confidence intervals" around estimates from probability samples, or ask whether a sample estimate differs in a statistically significant manner from some assumed population value. In the case of confidence intervals, one can obtain a formal assessment of how much "wiggle" there is likely to be in one's sample estimates. In the case of significance tests, one can reach a decision about whether a sample statistic (e.g., a mean reading score) differs from some assumed value in the population. For example, if the mean reading score from a random sample of students differs from some national norm, one can determine if the disparities represent statistically significant differences.

A second kind of chance factor stems from the process by which experimental subjects may be assigned to experimental and control groups. For example, it may turn out that the assignment process yields an experimental group that on the average contains brighter students than the control group. As suggested earlier, this may confound any genuine treatment effects with a priori differences between experimentals and controls; here the impact of some positive treatment such as self-paced instruction will be artifactually enhanced because the experimentals were already performing better than the controls.

Much as in the case of random sampling, when the assignment is undertaken with probability procedures, the role of chance factors can be taken into account. In particular, it is possible to determine the likelihood that outcome differences between experimentals and controls are statistically significant. If the disparities are statistically significant, chance (through the assignment process) is eliminated as an explanation, and the evaluator can then begin making substantive sense

of the results. If the process by which some units get the treatment and others do not is not a random process, one risks a "sample selection" bias that cannot be assessed with statistical inference. It is also possible to place confidence intervals around estimates of the treatment effect(s) which are usually couched as differences between the means on one or more outcome measures when the experimentals are compared to the controls. Again, an estimate of the "wiggle" is produced; in this case the "wiggle" refers to estimates of the experimental-control outcome differences.

A third kind of chance factor has nothing to do with research design interventions undertaken by the researcher (i.e., random sampling or random assignment). Rather, it surfaces even if a given population of interest is studied and no assignment process is undertaken. In brief, if one proceeds with the assumption that whatever the educational processes at work, there will be forces that have no systematic impact on outcomes of interest. Typically, these are viewed as a large number of small, random perturbations that on the average cancel out. For example, performance on a reading test may be affected by a child's mood, the amount of sleep gotten on the previous night, the quality of the morning's breakfast, a recent quarrel with a sibling, distractions in the room where the test is taken, anxiety about the test's consequences and the like. While these each introduce small amounts of variation in a child's performance, their aggregate impact is taken to be zero on the average (i.e., their expected value is zero). Yet since the aggregate impact is only zero on the average, the performance of particular students on particular days will be altered. Thus, there will be chance variation in performance that needs to be taken into account. And as before one can apply tests of statistical inference or confidence intervals. One can still ask, for example, if some observed difference between experimentals and control is larger than might be expected from these chance factors and/or estimate the "wiggle" in experimental-control disparities.

It is important to stress the statistical conclusion validity speaks to the quality of inferential methods applied and not to whether some result is statistically significant. Statistical conclusion validity

may be high or low independent of judgments about statistical significance. (For a more thorough discussion of these and other issues of statistical inference in evaluation research see Berk & Brewer, 1978.)

In summary, evaluation research involves a number of questions linked to different stages in a program's life history. Appropriate evaluation tools must be selected with such stages in mind and, in addition, against the four criteria just discussed. In other words, at each stage one or more policy relevant questions may be raised. Then, evaluation procedures should be selected with an eye to their relative strengths and weaknesses with respect to: measurement quality, an ability to weigh alternative causal explanations, the prospects for generalizing, and their capabilities for assessing the role of chance.

In the next few pages the general issues just raised will be addressed in more depth. However, before proceeding, it is also important to note that in the "real world" of evaluation research, even when an ideal marriage is made between the evaluation questions being posed and the empirical techniques to be employed, practical constraints may intervene. That is, questions of cost, timeliness, political feasibility and other difficulties may prevent the ideal from being realized. This in turn will require the development of a "second best" evaluation package (or even third best), more attuned to what is possible in practice. On the other hand, practical constraints do not in any way validate a dismissal of technical concerns; if anything, technical concerns become even more salient whan less desirable evaluation procedures are employed.

POLICY ISSUES AND CORRESPONDING EVALUATION STRATEGIES

This section will consider each of the major policy and program questions in turn and will identify the appropriate evaluation research strategies that are best fitted to provide answers to each of the policy questions.

I. Policy Formation and Program Design Issues

We first consider policy questions that arise in the policy formation and program design stage. Policy changes presumably arise out of dissatisfaction with existing policy, existing programs, or out of the realization that a problem exists for which a new policy may be an appropriate policy remedy. The information needed by policy makers and administrators is that which would make the policy and accompanying programs relevant to the problem as identified and efficacious in providing at least relief from some of the problem's burdens. It is important to stress that defining a "social problem" is ultimately a political process whose outcomes are not simply an assessment of available information. Thus, while it would be hard to argue against providing the best possible data on potential areas of need, there is no necessary correspondence between patterns in those data and what eventually surfaces as a subject of concern. (See for example, Berk & Rossi, 1976, for a more thorough discussion.)

As indicated earlier, we do not mean to imply by the organization of this section that policy makers always ask each of the questions raised in the order shown. The questions are arranged from the more general to the more specific, but that is an order we have imposed and is not intended to be a description of typical sequences or even a description of any sequence. Indeed, often enough, for example, research that uncovers the extent and depth of a social problem may spark the need for policy change, rather than vice versa, as may appear to be implied in this section.

Where is the Problem and How Much? The Needs Assessment Question

These are questions that seek to understand the distribution and extent of a given problem. Thus, it is one thing to recognize that some children are learning at a rate that is too slow to allow them to leave elementary schools sufficiently prepared for high school, and it is quite another to know that this problem is more characteristic of poor children and of minorities and more frequently encountered in inner city schools. It does not take more than a few instances of slow learning to document that a learning problem exists. To provide

sufficient information about the numbers of children who are in that deprived condition and to identify specific school systems with heavy concentrations of such children is quite another task. Similar questions arise with respect to other conditions that constitute the recognized social problems of our times, e.g., the distribution of quality medical care, adequate housing, and so on.

There are numerous examples of needs assessments that might be cited. Indeed, the monthly measurement of the labor force is perhaps the most extensive effort at needs assessment, providing a monthly estimate of unemployment and its distribution structurally and areally. The Office of Economic Opportunities 1968 Survey of Economic Opportunity was designed to provide a finer grained assessment of the extent and distribution of poverty in urban areas than was available through the decennial Census. The Coleman et al. (1967) report of educational opportunity was mandated by Congress to provide an assessment of how educational services and facilities were distributed among the poor.

The number of local needs assessments covering single municipalities, towns or counties done every year must now mount to the thousands. The Community Mental Health legislation calls for such researches to be undertaken periodically. Social impact statements to be prepared in advance of large scale alterations in the environment often call for estimates of the numbers of persons or households to be affected or to be served. The quality of such local assessments varies widely and is most likely on the average quite poor. The problem in attaining high quality needs assessments lies in the fact that the measurement of social problems of the more subtle variety (e.g., mental health) is quite difficult and the surveying methods that need to be employed are often beyond the reach of the talents and funds available.

It should be noted that the research effort involved in providing answers to the needs assessment question can be as inexpensive as copying relevant information from printed volumes of the U.S. Census to several years effort involving the design, fielding and analysis of a large

scale sample survey. Moreover, needs assessments do not have to be undertaken solely with quantitative techniques. Ethnographic research may also be instructive, especially in getting detailed knowledge of the specific nature of the needs in question. For example, the development of vocational training programs in secondary schools should respond to an understanding of precisely what sorts of job related skills are lacking in some target population. Perhaps the real need has more to do with how one finds a job commensurate with one's abilities than with an overall lack of skills per se (Liebow, 1964). On the other hand, when the time comes to assess the extent of the problem, there is no substitute for formal quantitative procedures. Stated a bit crudely, ethnographic procedures are likely to be especially effective in determining the nature of the need. Quantitative procedures are, however, essential when the extent of the need is considered.

While needs assessment research is ordinarily undertaken for the primary mission of developing accurate estimates of the amounts and distribution of a given problem, and hence is intended to be descriptive, often enough such research also can yield some understanding of the processes involved in the generation of the problem in question. For example, a search for information on how many high school students study a non-English language may bring to light the fact that many schools do not offer such courses and hence that part of the problem is that universally available opportunities to learn foreign languages may not exist. Or, the fact that many primary school children of low socioeconomic backgrounds appear to be tired and listless in class may be associated with a finding that few such children ate anything at all for breakfast before coming to school. A program that provided in-school breakfast feeding of poor children may be suggested by the findings of this needs assessment.

Particularly important for uncovering process information of this sort are carefully and sensitively conducted ethnographic studies. Thus ethnographic studies of disciplinary problems within high schools

may be able to point out promising leads as to why some schools have fewer disciplinary problems than others in addition to providing some indication of how widespread are problems associated with discipline. The findings on why schools differ might serve to suggest useful ways in which new programs could be designed that would help to bring all schools into line with those that are currently better at handling discipline issues.

Can We Do Anything About a Problem? Policy Oriented General Research

Knowing a lot about the distribution and extent of a problem does not by itself lead automatically to programs that can help to ameliorate that problem. In order to design programs we have to call upon two sorts of knowledge: first of all, basic social science understanding of a problem helps to point out the leverage points that may be used to change the distribution and extent of a problem. Secondly, we need to know something about the institutional arrangements that are implicated in a problem so that workable policies and programs can be designed. For example, our basic understanding of how students learn might suggest that lengthening the school day would be an effective way of increasing the rate of learning of certain skills. However, in constructing a program, we would have to take into account the fact that the lengthening of the school day is a matter that would concern teachers and their organizations as well as factors involving the capacity of schools to do so, other persons involved, including parents and school infra-structure personnel, etc.

Another example may help to illustrate how complex are the problems that arise in the design of appropriate programs. To know that there exist learning disabilities among school children by itself does not suggest what would be an appropriate policy response. To construct a policy response that has a chance to ameliorate educational problems typically means that there are some valid theories about how such problems arise and/or how such problems could be reduced. To pursue the learning disabilities question further, an appropriate set of knowledge

useful to policy formation would be theories that link learning disabilities to school experiences. Note that it is not crucial that learning disabilities be created by school experiences but only that school experiences influence to some appreciable degree the development of learning disabilities. There is little that policy can do (at least in the short run) about those "causes" of learning disabilities which have their roots in factors that are traditionally thought to be outside the sphere of policy relevance. Hence, knowledge about the role of family relationships in learning disabilities is not policy relevant (at present) because it concerns causes with which the policy sphere has traditionally not concerned itself. In contrast, research and knowledge dealing with the effects of schools, teachers, educational facilities and the like are currently policy relevant because social policy has been directed towards changing schools, the supply of educational facilities, the quality of teachers and similar issues.

This conception of policy relevant research is one that causes considerable misunderstanding concerning the relationships between basic and applied social research. A policy oriented research is one that tries to model how policy changes can affect the phenomenon in question. Knowledge about the phenomenon per se--the province of basic disciplinary concerns--may be important to understand how policy might be changed to alter the course of the social problem in question, but such basic research often does not. For example, laboratory studies of learning processes or of the development of aggression in persons may not be at all useful to educational policy makers or to criminal justice officials. Perhaps the clearest way to put the difference is that policy oriented and basic research are not contradictory or in conflict with each other but that in addition to understanding processes, policy oriented research must also be concerned with the connections between the phenomenon and how policies and programs may affect the phenomenon in question.

In addition, to construct a program that is likely to be adopted by an organization, we need to have intimate knowledge of what motivates such systems to change and adopt new procedures. Like other large scale

organizations, schools, factories, social agencies and the like are resistant to change, especially when the changes do not involve corresponding changes in reward systems. For example, an educational program that is likely to work is one that provides positive incentives for school systems and individual teachers to support and adopt the changes in learning practices embodied in the program.

Inadequate attention to the organizational contexts of programs is one of the more frequent sources of program implementation failure. Mandating that a particular program be delivered by an agency that is insufficiently motivated to do so, is poorly prepared to do so, and has personnel that do not have the skills to do so is a sure recipe for degraded and weakened interventions. Indeed, sometimes no programs at all are delivered under such circumstances (Rossi, 1978).

Answers to the question "Can we do anything about the problem?" can come from a variety of sources. Existing basic research efforts (whatever the method) aimed at understanding general educational processes are one source, although mastering this diverse technical literature is often difficult. Commissioned review papers may be an easy way to bring together in a pre-digested form the set of relevant existing basic research findings.

It should be noted that basic research is often not useful to policy needs because policy relevant concerns have not been directly addressed in the research. For example, studies of children who are disciplinary problems in school may stress understanding the links between the family situations of the children and their behavior. But, for policy and programmatic purposes, it would be considerably more useful if the researchers had spent their time studying how disciplinary systems within schools affect the rates at which disciplinary problems appeared within schools. Policy mutable variables (those that can be changed by policy) often tend to be slighted in basic research since policy is ordinarily only a small contributor to the total causal system that gives rise to a problem.

General research consciously linked to the role that schools and the educational system generally play in learning and other behavior

may be the best answer to policy needs. Such research would pay special attention to policy mutable conditions of individual behavior. Policy relevant general research may take a variety of forms, ranging all the way from systematic observational studies of school children to carefully controlled randomized experiments that systematically vary the policy relevant experiences of children. Without slighting basic research support, it should be emphasized that fostering such policy relevant general research needs special grant and contract research programs with review personnel that are familiar with what is policy relevant. It should be noted further that policy relevant general research should be accomplished with greater care and with more careful attention to problems of internal, external and construct validity precisely because of the importance that such research may have in the formation of policy and the design of programs.

Will Some Particular Program Work?

When some particular program has been identified that appears to be sensible according to current basic knowledge in the field, then the next step is to see whether it is effective enough to be worth developing into a program. It is at this point that we recommend the use of randomized controlled experiments in which the candidate programs are tested. Such research is extremely powerful in detecting program effects, because randomly allocating persons (or other units, e.g., classes) to an experimental group (to which the tested program is administered) or to a control group (from whom the program is withheld) assures that all the factors that ordinarily affect the educational process in question are on the average distributed identically among those who receive the program compared to those who do not. Therefore, randomization on the average eliminates causal processes that may be confounded with the educational intervention and hence enormously enhances internal validity. That is, the problem of spurious interpretations can be quite effectively addressed.

We advocate the use of randomized experiments at this stage in the development of a program both because they are powerful, in the sense

used above, and also because a potentially useful program ought to have the best chance of working when administered in a program that is run by dedicated researchers. However, this commitment in no way undermines the complementary potential of ethnographic studies, particularly to document why a particular intervention succeeds or fails.

Developmental experiments should be conducted ordinarily on a relatively modest scale and are most useful to policy needs when they test a set of alternative programs that are intended to achieve the same effects. Thus, it would be more useful for an experiment to test several ways of ameliorating learning disabilities since the end result would be to provide information on several possibly equally attractive (a priori) methods of ameliorating that condition.

There are many good examples of the field testing through randomized experiments of promising programs. The five income maintenance experiments were devised to test under varying conditions the impact of negative income tax plans as substitutes for existing welfare programs (Kershaw & Fair, 1976; Rossi & Lyall, 1976). The Department of Labor tested the extension of unemployment benefit coverage to prisoners released from state prisons in a small randomized experiment conducted in Baltimore (Lenihan, 1976). Randomized experiments have also been used to test out national health insurance plans and direct cash subsidies for housing to poor families. At issue in most of the randomized experiments were whether the proposed programs would produce the effects intended and whether undesirable side-effects could be kept at a minimum. Thus the Department of Labor LIFE experiment (Lenihan, 1976) was designed to see whether released felons would be aided to adjust to civilian life through increased employment and lowered arrest rates.

Can Agencies Deliver an Effective Program? Field Testing the Program

Once an effective treatment has been isolated, the next question that can be raised is whether a program incorporating the treatment can be administered through government agencies. Implementation of programs is always somewhat an open question. Agencies are no different from other organizations in resisting changes that are unfamiliar and

perhaps threatening. Interventions that work well with dedicated researchers administering them often fail when left to less skillful and less dedicated persons, as one might find in federal, state or local agencies. Hence, it is necessary to test whether agencies can deliver interventions at the proper dosage levels and without significant distortions. Randomized controlled experiments, as described above, are again an extremely powerful tool and appropriately designed randomized experiments should compare several modes of delivery to be tested.

Such field testing has been undertaken in a systematic way in a number of human services areas. For example, the Department of Housing and Urban Development commissioned ten cities to carry out demonstrations of housing allowance programs in order to test how best to administer such programs, leaving it to each city housing agency to set up its housing allowance program within the broad limits of specified payment levels and rules for client eligibility. Following up on the LIFE experiment noted above, the TARP experiments funded by the Department of Labor provide another example: Two states were chosen to run a program which provided eligibility for unemployment benefits to persons released from those states' prisons. Each state ran the program as a randomized experiment with payments provided through the Employment Security Agency of each state (Rossi, Berk, & Lenihan, 1980).

A special role at this stage can be played by process research activities which employ specially sensitive and observant researchers in close contact with the field testing sites. Observations collected by such researchers may be extremely useful in understanding the specific concrete processes that either impede or facilitate the working of the program. Again, ethnographic accounts can be extremely instructive in addressing the "whys."

II. Accountability Evaluation

Once a program has been enacted and is functioning, one of the main questions that is asked concerns whether or not the program is in place appropriately. Here the issues are not so much whether the

program is achieving its intended effects, but whether the program is simply running in ways that are appropriate and whether problems have arisen in the field to which answers need to be given. Programs often have to be fine-tuned in the first few years or so of operation.

Is the Program Reaching the Appropriate Schools and Children?

Achieving appropriate coverage of beneficiaries is often problematic. Sometimes a program may simply be designed inadvertently so as to be unable to reach and serve significant portions of the total intended beneficiary population. For example, a program designed to provide food subsidies to children who spend their days in child care facilities, may fail to reach a large proportion of such children if regulations exclude child care facilities that are serving fewer than five children. A very large proportion of children who are cared for during the day outside their own households are cared for by women who take in a few children into their homes (Abt Associates, 1979).

Experience with social programs over the past two decades has shown that there are few, if any, programs that achieve full coverage or near full coverage of intended beneficiaries, especially where coverage depends on positive actions on the part of beneficiaries. Thus not all persons who are eligible for Social Security payments actually apply for them, estimates range up to 15% of all eligible beneficiaries. Still others may not be reached because facilities for delivering the services involved are not accessible to them. And so on.

Although a thorough needs assessment of child care problems would have brought to light the fact that so large a proportion of child care was furnished by small scale vendors, and hence should have been taken into account in drawing up administrative regulations, such might not have been the case. In addition, patterns of the problem might change over time, sometimes in response to the existence of a program. Hence there is some need to review from time to time how many of the intended beneficiaries are being covered by a program.

There is also another side to the coverage problem. Programs may cover and extend benefits to persons or organizations that were not intended to be served. Such unwanted coverage may be impossible to

avoid because of the ways in which the program is delivered. For example, while Sesame Street was designed primarily to reach disadvantaged children, it also turned out to be attractive to advantaged children and many adults. There is no way that one can keep people from viewing a television program once broadcast (nor is it entirely desirable to do so in this case) and hence a successful TV program designed to reach some children may reach them but also many others (Cook, et al., 1975).

While the unwanted viewers of Sesame Street are reached at no additional costs, there are times when the "unwanted" coverage may turn out to severely drain program resources. For example, while Congress may have desired to provide educational experiences to returning veterans through the GI Bill and its successors, it was not clear whether Congress had in mind the subsidization of the many new proprietary educational enterprises that came into being primarily to supply "vocational" education to eligible veterans. Or, in the case of the bilingual education program, many primarily English speaking children were found to be program beneficiaries.

Studies designed to measure coverage are similar in principle to those discussed under "Needs Assessment" studies earlier. In addition, overcoverage may be studied as a problem through program administrative records. Undercoverage, however, often involves in many cases commissioning special surveys.

Are Appropriate Benefits Being Delivered? Program Integrity Research

When program services depend heavily on the ability of many agencies to recruit and train appropriate personnel or to retrain existing personnel or to undertake significant changes in standard operating procedures, it is sometimes problematic whether a program will always manage to deliver to beneficiaries what had been intended. For many reasons the issue of program integrity often becomes a critical one that may require additional fine-tuning of basic legislation or of administrative regulations.

Several examples may highlight the importance of this issue of educational programs. While funds may be provided for school systems to upgrade their audio-visual equipment, and schools may purchase them,

it is often the case that such equipment goes unused either because there are no persons trained to use the equipment or because audio-visual materials are not available (Rossi & Biddle, 1966). Or a new curriculum may be designed and made available to schools, but few schools are able to use the curriculum because teachers find the curriculum too difficult to use.

In other cases, the right services are being delivered but at a level that is too low to make a significant impact on beneficiaries. Thus a supplementary reading instruction program that means an additional forty minutes per week of reading instruction may not be delivered at sufficient strength and quantity to make any difference in reading progress.

Evaluation research designed to measure what is being delivered may be designed easily or may involve measurement problems of considerable complexity. Thus it may be very easy to learn from schools how many hours per week their audio-visual equipment is used, but very difficult to learn what is precisely going on inside a classroom when teachers attempt to use a new teaching method, where the program implies changes in teaching methods, classroom organization or other services that are highly dependent on persons for delivery. Measurement that would require direct observation of classroom activity may turn out to be very expensive to implement on a large scale.

Often for purposes of fine-tuning a program, it may not be necessary to proceed on a mass scale in doing research. Thus, it may not matter whether a particular problem in implementing a program occurs frequently or infrequently, since if it occurs at all it is not desirable. Hence for program fine-tuning small scale qualitative observational studies may be most fruitful.

Programs that depend heavily on personnel for delivery and/or which involve complicated programs and/or which call for individualized treatments for beneficiaries are especially good candidates for careful and sensitive fine-tuning research. Each of the characteristics enumerated in the previous sentence are ones that facilitate difficulties in appropriate implementation. In effect, this statement means that

personalized human services that are complicated are problematic in motivating personnel to deliver the services appropriately and skillfully.

Are Program Funds Being Used Appropriately? Fiscal Accountability

The accounting profession has been in operation considerably longer than has program evaluation. Hence procedures for determining whether or not program funds have been used responsibly and as intended are well established and hence are not problematic. However, it should be borne in mind that fiscal accountability measurements cannot substitute for the studies mentioned above. The fact that funds appear to be used as intended in an accounting sense may not mean that program services are being delivered as intended, in the sense discussed above. The conventional accounting categories used in a fiscal audit are ordinarily sufficient to detect, say fraudulent expenditure patterns, but may be insufficiently sensitive to detect whether services are being delivered in the requisite level of substantive integrity.

It is also important to keep in mind that the definition of costs under accounting principles differs from the definition of costs used by economists. For accountants, a cost reflects conventional bookkeeping entries such as out-of-pocket expenses, historical costs (i.e., what the purchase price of some item was), depreciation and the like. Basically, accountants focus on the value of current stocks of capital goods and inventories of products coupled with "cash flow" concerns. When the question is whether program funds are being appropriately spent, the accountant's definition will suffice. However, economists stress opportunity costs defined in terms of what is given up when resources are allocated to particular purposes. More specifically, opportunity costs reflect the next best use to which the resources could be put. For example, the opportunity cost of raising teachers' salaries by 10% may be the necessity of foregoing the purchase of a new set of textbooks. While opportunity costs may not be especially important from a cost-accounting point of view, opportunity costs become critical when cost-effectiveness or benefit-cost analyses of programs are undertaken. We will have more to say about these issues later.

III. Program Assessment Evaluation

The evaluation tasks discussed under accountability studies are directed mainly to questions dealing with how well a program is running. Whether or not a program is effective is a different issue, to which answers are not easily provided. Essentially, the question asks whether or not a program achieves its goals over and above what would be expected about the program.

Many evaluators consider that the effectiveness question is quintessentially evaluation. Indeed, there is some justification for that position since effectiveness assessment is certainly more difficult to accomplish, requiring higher levels of skills and ingenuity than any of the previously discussed evaluation activities. However, there is not justification for interpreting every evaluation task as calling for effectiveness assessments, as apparently some evaluators have done in the past, aided in their misinterpretation by imprecise requests for help from policy makers and administrators.

Can Effectiveness of a Program be Estimated? The Evaluability Question

A program that has gone through the stages described earlier in this chapter should provide few obstacles to evaluation for effectiveness in accomplishing its goals. But there are many human services programs that present problems for effectiveness studies because one or more of several criteria for evaluation are absent. Perhaps the most important criterion, one which is frequently absent, is the lack of well formulated goals or objectives for the program. For example, a program that is designed to raise the level of learning among certain groups of school children through the provision of per capita payments to schools for the purpose is not evaluable for its effectiveness without considerable further specification of goals. Raising the level of learning as a goal has to be specified further to indicate what is meant by "levels" and the kinds of learning achievements that are deemed relevant.

A second criteria is that the program in question be well specified. Thus a program that is designed to make social work agencies be more

effective by encouraging innovations is also not evaluable as far as effectiveness is concerned. First, the goals are not very well specified, but neither are the means for reaching goals. Innovation as a means of reaching a goal is not a method, but a way of proceeding. Anything new is an innovation and hence such a program may be encouraging the temporary adoption of a wide variety of specific techniques and is likely to vary widely from site to site.

Finally, a program is evaluable from an effectiveness point of view only if it is possible to estimate in some way what is the expected state of beneficiaries in the absence of the program. As we will discuss below, the critical hurdle in effectiveness studies is to develop comparisons between beneficiaries that experience a program with those who have not had such experiences. Hence a program that is universal in its coverage and that has been going on for some period of time cannot be evaluated for effectiveness. For example, we cannot evaluate the effectiveness of the public school systems in the United States, because it is not possible to make observations on Americans, cities, towns, counties and states that do not (or recently have not had) public school systems.

Finally effectiveness evaluations are the most difficult evaluation tasks undertaken by evaluators, requiring the most highly trained personnel for their undertaking, and considerable sums of money for data collection and analysis. Few evaluation units have the expertise and experience to design and/or carry out effectiveness evaluations. Especially rare are such capabilities on the state and local levels.

This discussion of effectiveness evaluability is raised here because we believe that often evaluators are asked to undertake tasks that are impossible or close to impossible. Thus it is not sensible for policy makers or program managers to call for effectiveness evaluation to be undertaken by all state and local evaluation units, at least at this stage in the development of state and local capabilities. Nor does it make much sense to undertake large scale evaluations of programs that have no nation-wide uniform goals but are locally defined.

Hence the evaluation of Title I or of Head Start and similar programs should not be undertaken or called for lightly, if at all.

Techniques have been developed (Wholey, 1977) to determine whether or not a program is evaluable in the senses discussed above. Congress and other decision makers may want to commission such studies as a first step rather than to assume that all programs can be evaluated.

Finally, it may be worth mentioning in passing that questions of evaluability have in the past been used to justify "goal-free" evaluation methods (e.g., Scriven, 1972; Deutscher, 1977). The goal-free advocates have contended that since many of a program's aims evolve over time, the "hypothetico-deductive" approach to impact assessment (Heilman, 1980) is at best incomplete and at worst misleading. In our view, impact assessment necessarily requires some set of program goals although whether they are stated in advance and/or evolve over time does have important implications for one's research procedures. In particular, evolving goals require far more flexible research designs (and researchers). In other words, there cannot be such a thing as a "goal-free" impact assessment. At the same time, we have stressed above that there are other important dimensions to the evaluation enterprise in which goals are far less central. For example, a sensitive monitoring of program activities can proceed productively without any consideration of ultimate goals. Thus, goal-free evaluation approaches can be extremely useful as long as the questions they can address are clearly understood.

Did the Program Work? The Effectiveness Question

As discussed above, any assessment of whether or not a program "worked" necessarily assumes that it is known what the program was supposed to accomplish. For a variety of reasons, enabling legislation establishing programs may appear to set relatively vague goals or objectives for the program and it is necessary during the "design phase" (as discussed above) to develop specific goals. Goals for such general programs may be developed by program administrators through consideration of social science theory, past research and/or studies of the problem

that the program is supposed to ameliorate. Thus Title I was designed to enrich the educational experiences of disadvantaged children through providing special funds to state and local school systems that have relatively large proportions of disadvantaged children on their rolls. However, in order to accomplish this general (and too general) objective, it was necessary in local school systems to develop specific programs with their own goals. Thus some goals or sets of objectives may be developed as a program goes along (Chen & Rossi, 1980).

However goals may be established, the important point is that it is not possible to determine whether a program worked without developing a limited and specific set of criteria for establishing the condition of "having worked." For example, it would not have been possible to develop an assessment of whether Sesame Street "worked" without having decided that its goals were to foster reading and number handling skills. Whether these goals existed before the program was designed or whether they emerged after the program was in operation is less important for our purposes than the fact that such goals existed.

Programs rarely succeed or fail in absolute terms. Success or failure is always relative to some benchmark. Hence an answer to "Did the program work?" requires a consideration of "Compared to what?"

The development of appropriate comparisons can proceed along at least three dimensions: comparisons across different subjects, comparisons across different settings and comparisons across different times. In the first instance, one might compare the performance of two sets of students in a given class in a given classroom period. In the second instance, one might compare the performance of the same set of students in two different classroom settings (necessarily at two different points in time). In the third instance, one might compare the same students in the same classroom, but at different points in time.

As Figure 1.1 indicates, it is also possible to mix and match these three fundamental dimensions to develop a wide variety of comparison groups. For example, comparison group 2 (C_2)* varies both the subjects

*We have used the term "comparison group" as a general term to be distinguished from the term "control group." Control groups are comparison groups that have been constructed by random assignment.

Figure 1.1

A TYPOLOGY FOR COMPARISON GROUPS

	Same Subjects		Different Subjects	
	Same Setting	Different Setting	Same Setting	Different Setting
Same Time	[a]	[a]	C_1	C_2
Different Time	C_3	C_4	C_5	C_6

[a]These two boxes, while logically possible, lead to comparison groups which make no sense substantively in this context.

and the setting although the time is the same. Or, comparison group 6 (C_6) varies subjects, the setting and the time. However, with each added dimension by which one or more comparison groups differ from the experimental group, the number of threats to internal validity necessarily increases. For example, the use of comparison group 4 (different setting and different time period) requires that assessment of program impact simultaneously take into account possible confounding factors associated with such things as differences in student background and motivation and such things as the "reactive" potential of different classroom environments. This in turn requires either an extensive data collection effort to obtain measures on these confounding factors coupled with the application of appropriate statistical adjustments (e.g., multiple regression analysis), or the use of randomization and thus, true control groups. Randomization, of course, will on the average eliminate confounding influences in the analysis of impact. On grounds of analytic simplicity alone, it is easy to see why so many expositions of impact assessment strongly favor research designs based on random assignment. In addition, it cannot be overemphasized that appropriate statistical adjustments (in the absence of randomization) through multivariate statistical techniques require a number of assumptions that are almost impossible to fully meet in practice.* For example, it is essential that measures of all confounding influences be included in a formal model of the program's impact, that their mathematical relationship to the outcome be properly specified (e.g., a linear additive form versus a multiplicative form), and that the confounding influences be measured <u>without error</u>! Should any of these requirements be violated, one risks serious bias in any estimates of program impact.

At the same time, however, random assignment is often impractical or even impossible. And even when random assignment is feasible, its advantages rest on randomly assigning a relatively <u>large</u> number of subjects. To randomly assign only two schools to the experimental group

*There are some research designs which while not based on random assignment, do readily allow for unbiased estimates of treatment effects through multivariate statistical adjustments. See, for example, Barnow, Cain and Goldberger (1980).

and only two schools to the control group, for example, will not allow on the average equivalences between experimentals and controls to materialize. Consequently, one is often forced to attempt statistical adjustments for initial differences between experimental and comparison subjects.

The use of multivariate statistical adjustments raises a host of questions that cannot be addressed in detail here. Suffice to say that despite the views of some that anything that can go wrong, will go wrong, extensive practical experience suggests a more optimistic conclusion. Quite often, useful and reasonably accurate estimates of program effects can be obtained despite modest violations of the required statistical assumptions. Moreover, available statistical technology is evolving rapidly and many earlier problems now have feasible solutions, at least in principle. (For a review of some recent statistical developments in the context of criminal justice evaluation, see Berk, 1980.)

To consider the usefulness of assessments not resting on random assignment, consider a recent evaluation (Robertson, 1980) of the effectiveness of driver education programs in reducing accidents among 16 to 18 year olds. The evaluator took advantage of the fact that the Connecticut legislature decided not to subsidize such programs within local school systems. In response to this move, some school districts dropped driver education out of their high school curriculum and some retained it. Two sets of comparisons were possible: accident rates for persons of the appropriate age range in the districts that dropped the program were computed before and after the program was dropped and accident rates for the same age groups in the districts that retained driver education compared to the accident rates in districts that dropped the driver education program. It was found that the accident rates significantly dropped in those districts that dropped the program, a finding that led to the interpretation that the program increased accidents because young people were led to obtain licenses earlier than otherwise.

It is sometimes possible to either enhance or partially bypass comparison group problems by resorting to some set of external criteria

as a baseline. For example, it is common in studies of desegregation or affirmative action programs to apply various measures of equity as a "comparison group" (Baldus & Cole, 1977). Thus, an assessment of whether schools in black neighborhoods are being funded at comparable levels to schools in white neighborhoods, might apply the criterion that disparities in excess of plus or minus 5% in per pupil expenditures indicate inequity and hence failure (Berk & Hartman, 1972). However, the use of such external baselines by themselves still leave open the question of causal inference. It may be difficult to determine if the program or some other set of factors produced the observed relationship between outcomes of interest and the external metric.

It is also important to understand that distinguishing between success and failure is not a clearcut decision since there are usually degrees of success or degrees of failure. While decision makers may have to make binary decisions whether, for example, to fund or not to fund, the evidence provided on effectiveness usually consists of statements of degree which then have to be translated into binary terms by the decision makers. Thus it may turn out that a program that succeeds in raising the average level of reading by half a year more than one would ordinarily expect to be reading gains, such a program may be less successful than one which has effectiveness estimates of a full year. This quantitative difference has to be translated into a qualitative difference when the decision to fund one rather than the other program comes into question.

In short, the construction of effectiveness evaluation studies is a task that requires a considerable amount of skill. Hence such effectiveness studies should be called for when there is sufficient reason to believe that the circumstances warrant such studies, as mentioned earlier in this chapter, and on whether or not capability is available in the unit responsible for the study.

Was the Program Worth It? The Economic Efficiency Question

Given a program of proven effectiveness, the next question one might reasonably raise is whether the opportunity costs of the programs are justified by the gains achieved.* Or the same question might be more narrowly raised in a comparative framework, is Program A more "efficient" than Program B, as alternative ways of achieving some particular goal?

The main problem in answering such questions centers around establishing a yardstick for such an assessment. For example, would it be useful to think in terms of dollars spent for units of achievement gained, in terms of students covered, or in terms of classes or schools that come under the program.

The simplest way of answering efficiency issues is to calculate cost effectiveness measures, dollars spent per unit of output. Thus in the case of the Sesame Street program, several cost effectiveness measures were computed:

--Dollars spent per child hour of viewing;

--Dollars spent per additional letter of the alphabet learned.

Note that the second measure implies knowing the effectiveness of the program, as established by an effectiveness evaluation.

The most complicated mode of answering the efficiency question is to conduct a full-fledged cost-benefit analysis in which all the costs and benefits are computed. Relatively few full-fledged cost-benefit analyses have been made of social programs because it is difficult to put all the costs and all the benefits into the same yardstick terms. In principle, it is possible to convert into dollars all the costs and benefits of a program. In practice it is rarely possible to do so without some disagreement on the valuation placed, say on learning an additional letter of the alphabet.

An additional problem with full-fledged benefit-cost analyses is that they must consider the long run consequences not only of the program, but the long run consequences of the next best alternative

*Recall that opportunity costs address the foregone benefits of the next best use of the resources in question (Thompson, 1980:65-74).

foregone. This immediately raises the question of "discounting:" the fact that resources invested today in some social program may produce consequences over a large number of succeeding years that have to be compared to the consequences from the next best alternative over a large number of succeeding years. For example, a vocational program in inner city high schools needs to address (among other things) the long run impact of students' earnings over their lifetimes. This in turn requires that the costs and benefits of the program and the next best alternative be phrased in terms of today's dollars. Without going into the arcane art of discounting, the problem is to figure out what a reasonable rate of return over the long run for current program investments and competing alternatives might be. And, one can obtain widely varying assessments depending on what rate of return is used (Thompson, 1980).

Evaluation in Evolution

The field of evaluation research is scarcely out of its infancy as a social scientific field of inquiry. The first large scale field experiments were started in the middle 60s. Concern for large scale national evaluations of programs also had their origins in the War on Poverty. The art of designing large scale implementation and monitoring studies is just now evolving. Concern with the validity statuses of qualitative research has just begun. And so on.

Perhaps what is most important as a developing theme is the importance of social science theory for evaluation. It has become increasingly obvious that social policy is almost a blind thrashing about for solutions. Guiding the formation of social policy through sensitive and innovative applications of general social science theory and empirical knowledge is beginning to occur more and more. This development is further enhanced by the increasingly held realization that errors in model specification are errors in theory. Hence there is no good policy without good understanding of the problem involved and of the role that policy can play. Nor is there any good evaluation without theoretical guidance in modelling policy effects.

REFERENCES

Abt Associates. Child care food program, Cambridge, MA: Abt Associates. 1979.

Baldus, D. C., & Cole, J. W. L. Quantitative proof of intentional discrimination. Evaluation Quarterly, 1977, 1(1).

Barnow, B. S., Cain, G. G,. & Goldberger, A. S. Issues in the analysis of selectivity bias. In E. W. Stromsdorfer & G. Farkus (Eds.), Evaluation Studies Review Annual, Volume 5, Beverly Hills: Sage Publications. 1980.

Becker, H. S. Problems of inference and proof in participant studies. American Journal of Sociology, 1958, 23(2).

Berk, R. A. Recent statistical developments with implications for evaluation of criminal justice programs. In M. Klein & K. Teilman, Handbook of Criminal Justice Evaluation, Beverly Hills: Sage Publications. 1980.

Berk, R. A., & Brewer, M. Feet of clay in hobnailed boots: an assessment of statistical inference in applied research. In T. D. Cook (ed.), Evaluation Studies Review Annual, Volume 3, Beverly Hills: Sage Publications, 1978.

Berk, R. A., & Hartman, A. Race and class differences in per pupil staffing expenditures in chicago elementary schools. Integrated Education, January, 1972.

Berk, R. A., & Rossi, P. H. Doing good or worse: evaluation research politically re-examined. Social Problems, 1976, 23(4).

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations make compensatory education look harmful. In J. Hellmuth (Ed.), Compensatory education: A national debate. New York: Brunner/Mazel, 1970.

Chen, H., & Rossi, P. H. The multi-goal, theory-driven approach to evaluation: A model linking basic and applied social science. Social Forces, 1980, 59(1).

Coleman, J., et. al. Equality of educational opportunity. Washington: Government Printing Office, 1967.

Cook, T., et. al. Sesame Street revisited. New York: Russell Sage. 1975.

Cook, T., & Campbell, D. Quasi-experimentation. Chicago: Rand McNally. 1979.

Cronbach, L.J. Five decades of controversy over mental testing, American Psychologist, 1975, 30(1).

Cronbach, L.J. Toward reform of program evaluation, Menlo Park, CA: Jossey-Bass, 1980.

Deutscher, I. Toward avoiding the goal trap in evaluation research. In F. Caro (Ed.), Readings in evaluation research, New York: Russell Sage, 1977.

Franke, R.H. The Hawthorne experiments: Review. American Sociological Review, 1979, 44(5).

Franke, E.M., & Kaul, J.D. The Hawthorne experiments: First statistical interpretation, American Sociological Review, 1978, 43(5).

Gramlich, E.M., & Koshel, P. Educational performance contracting, Washington: The Brookings Institution, 1975.

Guttentag, M. & Struening, E. (Eds.) Handbook of evaluation research (2 vols.). Beverly Hills: Sage Publications, 1975.

Heilman, J.G. Paradigmatic choices in evaluation methodology. Evaluation Review, 1980, 4(5).

Kershaw, D., & Fair, J. The New Jersey income maintenance experiment. New York: Academic, 1976.

Kish, L. Survey sampling. New York: Wiley, 1965.

Kmenta, J. Elements of econometrics. New York: MacMillan, 1971.

Lenihan, K. Opening the second gate. Washington: Government Printing Office, 1976.

Liebow, E. Tally's corner. Boston: Little-Brown, 1964.

Mathematica Policy Research. Jobs Corps evaluated. Princeton: Mathematica, 1980.

Maynard, R.A., & Murnane, R.J. The effects of the negative income tax on school performance. Journal of Human Resources, 1979, 14(4).

Mensh, I.N., & Henry, J. Direct observation and psychological tests in anthropological field work. American Anthropology, 1955, 55(6).

Robertson, L.S. Crash involvement of teenaged drivers when driver education is eliminated from high school. American Journal of Public Health, 1980, 70(6).

Rossi, P.H. Issues in the evaluation of human services delivery. Evaluation Quarterly, 1978, 2(4).

Rossi, P.H., Berk, R., & Lenihan, K. Money, work and crime. New York: Academic Press, 1980.

Rossi, P.H., & Biddle, B. The new media and education. Chicago: Aldine, 1966.

Rossi, P.H., Freeman, H., & Wright, S. Evaluation: A systematic approach. Beverly Hills: Sage Publications, 1979.

Rossi, P.H., & Lyall, K. Reforming public welfare, New York: Russell Sage, 1974.

Scriven, M. Prose and cons about goal-free evaluation. Evaluation Comment, 1972, 3(1).

Suchman, E. Evaluation research. New York: Russell Sage, 1967.

Sudman, S. Applied sampling. New York: Academic Press, 1976.

Thompson, M. Cost-benefit analysis. Beverly Hills: Sage Publications, 1980

Wardwell, W.L. Comment on Kaul and Franke. American Sociological Review, 1979, 44(5).

Weiss, C. Evaluation research. Englewood Cliffs, NJ: Prentice Hall, 1972.

Wholey, J.S. Evaluability assessment. In L. Rutman (Ed.) Evaluation research methods, Beverly Hills: Sage Publications, 1977.

Zelditch, M. Some methodological problems in field studies. American Journal of Sociology, 1962, 67(2).

CHAPTER 2

USES AND USERS OF EVALUATION

Marvin C. Alkin
University of California, Los Angeles

Introduction

Social scientists like to attribute rationality to the various activities conducted in social systems. Evaluation is one of those presumable rational activities. Indeed, the presumption holds that evaluation's rationality is attributed to its purposiveness--it serves a useful purpose. It may be safely said that all evaluations serve a purpose perceived to be useful to someone. An evaluation might be conducted simply to satisfy legislative requirements, or to satisfy the illusion that the organization is engaged in systematic self-evaluation. In some quarters, each of these is considered to be a useful purpose, or else they would most certainly not have been initiated.

Four types of situations have been outlined (Alkin, 1975) which do not aim evaluation at decision making: (1) window-dressing, (2) legal requirements, (3) public relations, and (4) professional prestige. Window dressing evaluations seek justification of decisions already made. Evaluations commissioned simply to comply with legal requriements often deliberately make evaluations solely a pro forma exercise. Some evaluations are commissioned simply as public relations gestures where the intent is to demonstrate the objectivity of decision makers. Professional prestige is designed to enhance individual reputations of the report commissioners as innovators.

However, it is reasonable to argue that judgment of evaluation rests not only on its serving a generally useful purpose, but on the extent it actually informs decision making--the most highly desired of potentially useful purposes. This point is well articulated by Weiss:

> the basic rationale for evaluation is that it provides information for action. Its primary justification is that it contributes to the rationalization of decision making. Although it can serve such other functions as knowledge-building and theory-testing, unless it gains serious hearing when program decisions are made, it fails in its major purpose. (1972; p. 318)

Unfortunately, there is wide-spread disagreement as to whom evaluation has provided "information for action," or more broadly, what constitutes use. Furthermore, it is not clear who may be considered "the users." In the remainder of this chapter, these two major issues will be addressed with respect to national evaluations in the Education Department: who are the users? and what constitutes use?

The Users

For evaluations to be used, there must be someone to make use of them. And, the nature of these appropriate users is dependent on the level at which evaluation takes place. Using the focus on the current study, national evaluations of education, as the basis for further analysis, the various categories of appropriate users will be examined. A second issue to be considered is the interplay between those who commission evaluation studies and other potential users.

Categories of users. While there are a number of ways to typify user groups, for the purposes of this chapter, the following category system will be employed: (1) Congress, (2) Education Department management, (3) program management, and (4) SEA and LEA management.

Although it may be fashionable to demean the use of evaluation research data by Congress, it is nonetheless clear that Congress is in fact a major user of evaluation research. Perhaps undue expectations of use prevail and the failure to consider a particular written evaluation report in toto as part of the re-funding decision might easily lead to the conclusion that evaluations are not used. Nevertheless, there is little question that the information from evaluation research filters into the system and is used over a period of time. A study consisting of twenty-six interviews with congressional staff members (Florio, 1980) provided evidence that legislative aides identified student achievement scores as the second most important types of information (next to the cost of the program). This evidence, along with the data on use presented in the U.S. Office of Education Annual Report, suggests strongly that Congress does, in fact, use the results of evaluation research. A limited sampling of the ways evaluation use takes place are: as input for making funding decisions with respect to the program; as input for changing the scope

of the program; in attitude formation about the program for potential future use.

> Examples: Changes in Title VII eligibility requirements were made in the Education Amendments of 1978 following from the OED "Evaluation of ESEA Title VII bilingual education program," 1974-1978; changes in Title I legislation in 1978 were made based in part on a variety of OED studies and in particular on a congressionally requested and focused study which was directed by Paul Hill of RAND Corporation.

A second category of user, the upper level management of the new Education Department, might appropriately employ evaluation information in a number of ways within the organizational structure. For example, evaluation research information might contribute to judgments about the quality of a program, the leadership being provided by program management, the scope of the program, or the appropriateness of the audience--to name but a few. It may also provide input for decisions about modifying required rules and regulations to obtain better conformity to the intended program strategies.

> Examples: A 1976 evaluation of Title IV of the Civil Rights Act was used as the basis for developing a set of twelve detailed recommendations for policy changes. These were reviewed by the Commissioner of Education, and final regulations incorporating the recommendations were issued in July 1978; a 1978 evaluation of the impact of ESEA Title VII had a major influence on decisions at the Department or Office level. The OE Commisioner noted during the 1979 Senate Appropriations Hearings: ". . .Senator, that study was very significantly related to some major moves we made. The first move we made was to change the director of that office. The second move we made was to call an internal audit that dealt with the staffing and program procedures in that office . . .. More than that, Secretary Califano established an internal tracking procedure in which we are to report quarterly . . . [on] bilingual education to demonstrate to him at least four times a year that our goals of increasing the number of deficient children in bilingual education was being met."

A large cadre of program managers within a government department are clearly the recipients and users of evaluation research data. Those who immediately come to mind manage the bulk of the programs charged with administering funds under a variety of program categories to state education agencies, local education agencies, universities, etc. Evaluation

research conducted for these people may provide an input for decisions about needed program review techniques or additional program monitoring which is required.

> Examples: The 1979 report on "OED Uses of Evaluation Activities" provides several illustrations of program management use of evaluation data: Evaluation of the National Diffusion Network and related studies on school improvement have influenced strategies to place more emphasis on the quality of implementation, fidelity of adoptions, and the impact of the programs on learners in the adopter sites. A second example: Studies have been completed for Upward Bound, Talent Search, and Special Services program for Disadvantaged Students. As a result, evaluation findings have been used in the writing and/or revision of regularities for the UB, TS, and SSDS programs so as to improve award procedures, overall program mangement, and monitoring and reporting procedures.

Another audience for nationally conducted evaluation research is the management personnel at state education agencies (SEA) and local education agencies (LEA)*. Results of national evaluations are provided to SEA and LEA officials, and it is anticipated that there will be SEA and LEA users. Evaluation research data from nationally conducted evaluations may provide input for rating state or local education agencies against other states or districts. Or, the evaluation research data might be used for making judgments with respect to the delivery system or instructional treatments used within the education agency.

> Example: Title I is generally provided as the prime example of SEA and LEA evaluation use. Employing a required set of evaluation procedures enables State and school districts to make more informed comparisons of program outcomes** Another example of SEA evaluation use is provided by the evaluation of state plans for career education. Based on the evaluation data, OED reports that "well over half of them [states] voluntarily provided revisions and/or additions in order to remedy weaknesses which had been pointed out to them."

---

*In this paper we have only considered "nationally conducted" evaluations, and not "nationally mandated" evaluations, such as the Title I evaluation conducted in the local school district. This orientation is in keeping with the Congressional Mandate for this study.

**It is not clear, however, whether mandated employment of the Title I TIERS evaluation reporting system is, in fact, an instance of "evaluation use." Must one first document that these procedures provide outcome data which districts use for decision making?

Information needs and the likelihood that evaluation will be used are clearly related to a particular user's organizational role. Thus, pre-specification of the anticipated evaluation user is imperative. In this section, the various possible user categories have been indicated.* But, the question remains--for a particular evaluation, who are the individuals/audiences who are most likely to use the evaluation?

Evaluation report commissioners and secondary users. It has been said that "he who pays the piper, calls the tune." And, House (1972) has written: "who sponsors and pays for the evaluation makes a critical difference in the evaluation findings" (p. 409). The argument may be extended to evaluation users: the source of the evaluation is an important factor in considering the potential user audiences and the extent of utilization.

The crucial distinction, however, is not those who fund the evaluation per se, or who pay the bill, but rather, those who hire the evaluator, set the agenda, direct his action, exercise control and oversight on the evaluation, etc. It is convenient to refer to these persons as "evaluation report commissioners." Evaluation report commissioners (ERCs) set the context of the evaluation in terms of what is considered acceptable content, what questions are to be answered, and even, which elements of the agency are to be subject to scrutiny. If the ERCs of a federal program are themselves program managers, then the evaluation will probably focus on procedural aspects, bases for program modification and improvement, deficiencies of SEAs and LEAs in implementing the federally funded program. Likewise, evaluators are likely to be less than candid in their report about major program characteristics, the quality of the program as a whole, or the quality of the program leadership. Partly, evaluators are, in fact, sensitive to "who is calling the tune" (with political overtones of future employment from the agency, etc.). Also, aside from the political aspects, there is simply the question of how the research agenda is limited: what are the allowable spheres, etc.

---

*In this paper we have only addressed organizationally related user groups. Other user groups include "the public" or special interest groups, for examples.

To further illustrate the extent to which the evaluation report commissioner is defined by "who controls the evaluation," and not simply by such matters as who formally lets the contract or pays the bills, consider the following example. The Title I Sustaining Effects Study, while formally contracted by NIE and housed as a contract within NIE, was none the less mandated by Congess. Moreover, the legislation allowed for frequent interim consultation with a congressional committee to discuss the progress and course of the evaluations, and provided for direct reporting of the evaluation results to Congress without approval by NIE program heads. Even though the formal contracting occurred within NIE, it is certainly appropriate to conceive of Congress (or at least the congressional committee involved) as the evaluation report commissioner. As a side-light, this particular evaluation study is perhaps most often cited as among those most useful to Congress.

Therefore, it is important to consider how national evaluations are commissioned in order to understand potential evaluation use. Within the Office of Education as previously constituted, evaluation activities were commissioned in at least three different locations. Some evaluations were initiated, supervised, etc. in the Office of the Assistant Secretary for Planning and Evaluation; the large bulk of the evaluations were commissioned through the Office of Evaluation and Dissemination; and some evaluation report commissioners were to be found among the various programs of the Office of Education. Within the National Institute of Education, evaluation reports were commissioned, by and large, at the program level.

Sometimes evaluations are intended by commissioners principally to have impact on secondary users. Department leadership might commission a report to provide needed evidence about a program. Knowing, for example, that congressional hearings will be shortly forthcoming on a particular program up for renewal, the commissioned evaluation would very likely aim primarily to provide information (hopefully positive) to congressional users. Moreover, secondary users might be found within the same department, as for example, when an evaluation is commissioned within a program already congressionally approved in order to provide data to department managers that might lead to modifications in the program. Many other examples could be provided of secondary user relationships.

However, not all evaluations have anticipated or possible secondary users. Sometimes the organizational role of the evaluation report commissioner and the spheres of inquiry in the evaluation may limit its utility primarily to ERCs. The development of an evaluation around questions primarily of interest to evaluation reports commissioners who had programs in the Office of Education may make the evaluation too limited in scope for Congress as a potential secondary user. Or, by the same token, program manager evaluation questions may be too aggregated and national in focus to be of use to SEAs and LEAs. Furthermore, different user categories may, by the nature of the information they see, impose quite different standards on appropriateness of evaluation methodology (e.g., a program manager may be quite satisfied with descriptive data as a source of information for program change while Congress or Department leadership are unwilling to accept data as convincing which does not employ an experimental design.) Indeed, these qualifications should not be construed as criticisms, for it may be impossible to develop evaluations that fully meet the information needs and acceptable standards of evaluation data of a variety of users.

The point is clear: the information required for users at one level may in fact preclude important data for other users. The Title I Evaluation and Reporting System (TIERS) provides an excellent example. In the attempt to develop an evaluation system to satisfy the information needs of all levels of users (from the classroom teacher to Congress), it may be that a system has been created that is not totally appropriate to any users. And, beyond the reporting system, it becomes even more obvious that the development of an evaluation report from TIERS data, of necessity, focuses on one level of aggregation which diminishes the report's value to users at other levels.

To sum up, in considering potential users for evaluations, it is important to examine who the evaluation report commissioners are, and the extent to which the evaluation has anticipated secondary users. The various management levels within the Office of Education as well as Congress, SEAs and LEAs offer a wide variety of evaluation report commissioners and anticipated secondary users.

Uses: What Constitutes Utilization?

There is no unified view of whether evaluations have impact on what constitutes evaluation use. One belief, well documented in the literature, contends that evaluations seldom influence program decision making--and countless articles reflecting this stance bemoan the unlikelihood that evaluation will ever break through the barriers and have real impact on programs. An alternative point of view, only recently expressed in the literature, reaches quite a different conclusion: that evaluations do already influence programs in important and useful ways.

The extent to which an individual propounds one or the other of these viewpoints is largely dependent on the definition of utilization that he/she employs. The group which decries the lack of use would undoubtedly employ a very restrictive definition which would require that a single intended user (typically the ERC) make a specific decision immediately following the receipt of an evaluation report and heavily (if not solely) based upon the findings of that report. Alternatively, it would be easy to find great evidence of utilization with a definition that encompasses any use of anything from the evaluation for purposes broadly conceived.

In our view, neither of these approaches to defining evaluation utilization is appropriate. Instead, they represent caricatures of definitions representing opposite ends of a continuum of views on utilization. Neither is workable; neither is realistic. In the remainder of this section, we will examine a variety of views on evaluation use from the literature and attempt to derive a more appropriate definition.

Other researchers have added substantively to the deliberations about a definition of evaluation utilization. Caplan, Morrison and Stambaugh (1975) have said: "utilization of knowledge . . . occurred when the respondent was familiar with relevant research and gave serious consideration to an attempt to apply that knowledge to some policy relevant issue" (p. VII). Furthermore, these authors have contributed the concepts of "instrumental" and "conceptual" utilization to the literature of the field. These concepts are further elaborated by references to instrumental use as where respondents cited and were able to document the specific way in which information was being used for decision making purposes. On the

other hand, conceptual use refers to those instances in which a policy maker's thinking was influenced by the evaluation or a policy maker planned to use information in the future.

There is general acceptance for including both instrumental and conceptual dimensions within a general definition of evaluation use. In an examination of research on evaluation utilization, Conner (1980) noted that five of the six major studies employed a definition of utilization which encompasses both instrumental and conceptual usage. (He concluded that with this broadened definition, usage generally was found to be high.) Knorr (1977) further extended this use category system by introducing the notion of "symbolic" use. Pelz (1978) draws the distinction between three different types of symbolic use: use as a substitute for a decision; use to legitimate a policy; and, use to support a predetermined position. The first of these types of symbolic use does not appear to be an actual use of the evaluation and moreover, has been discussed earlier in this chapter (as "window-dressing" and "public relations" evaluations). The latter two symbolic use types appear to involve a common theme--substantiating a previously made decision or current point of view.

Alkin, Daillak and White (1979), in their recent study, have attempted to isolate the essential components of utilization and present their definition of utilization in the form of a Guttman facet design sentence. The facets include: (1) the nature of the client (e.g., evaluation report commissioner); (2) the nature of the use (e.g., one of multiple influences); (3) the type of use (e.g., making a decision); (4) and the topic of use (e.g., continuance of a program component). The notions of identifying both primary and secondary users and of instrumental and conceptual uses appear to be encompassed within this definition.

More recently, Weiss (1980) has provided additional elaboration on a theme prevalent in the literature by discussing the extent to which research or evaluation information is utilized within systems primarily on very gradual bases, or, over long periods of time. The concept of "knowledge creep"--of incremental, temporarily gradual use of evaluation information--has also been discussed, to some extent, by Caplan et al. (1957), Patton et al. (1975), Alkin et al. (1974), and Alkin et al. (1979).

Drawing from these sources--the prevalent literature in the field--a definition of utilization is presented for purposes of this report. The definition is in the form of a simple matrix depicting instances of evaluation use. (See Table 2.1.) As seen in the table, the various categories of evaluation use are fairly obvious. Evaluation information may be used: to substantiate a prior decision, as input to a current decision, or as part of general attitude information.

The first and third of these categories are fairly self-descriptive. The second requires additional clarification. Three subsets have been described for the second category--"input to a current decision." First, evaluation information may be the primary basis for making a decision. It probably is quite naive to expect that policy decisions will be made based solely on evaluations; however, there are instances in which evaluation provides the primary information basis for policy action. Or, evaluations become mingled with other data input--personal views of decision makers, judgments of political difficulty, etc.--to determine a policy decision. A third subcategory is perhaps difficult to distinguish from the second (and indeed, may be the same)--evaluation as one thread in the fabric of cumulative inputs over time. In this instance, perhaps there have been prior evaluation reports in prior years. Possibly, the evaluation information received in the current year is just that piece of additional information which stimulates some incremental policy change. This is best described by one of the participants in a conference on evaluation use in federal agencies.

> The way in which evaluation contributes is through the Chinese water torture approach. Each study adds a little bit of information, only a little bit, and after a good many studies you begin to feel it a little more and a little more . . .. So in terms of the impact of evaluation on broad program direction and policies it has that kind of cumulative effect, and those who ask which study led to the termination of a particular program, just don't understand either decision making or evalation. (Chelimsky, 1976)

It may well be that this subcategory is most pervasive and includes most of the instances of documented evaluation use for policy decision making.

Table 2.1

EVALUATION USE MATRIX

	Substantiate a previous decision or point of view	Primary basis for a decision	One of multiple current inputs for a decision	One of multiple cumulative (temporal) inputs for a decision	Change attitudes
Evaluation Report Commissioner					
Intended Secondary User					

Instead of attempting to provide examples for each of the cells of the evaluation use matrix, several examples should provide sufficient elaboration to make the definition workable:

> Example--Congress, as an Evaluation Report Commissioner, uses Title I sustaining effects data along with the results of Title I evaluations in previous years, the views of constituents, and other testimony to refund and make changes in the Title I program.
>
> Example--The findings of OE studies to identify effective projects in compensatory education was the primary basis for the decision as to which of these projects is to be included within the National Diffusion Network.
>
> Example--An evaluation of Title VII provided data which was the primary input for the decision by Congress as an influenced secondary user to change the Title VII eligibility requirements.

Summary

In this paper various categories of users have been described along with the distinction between evaluation report commissioners and secondary users. Furthermore, a matrix has been presented explaining conditions constituting evaluation use. A conceptualization of evaluation users and evaluation use such as this raises a host of procedural and politically related issues. It will be important to recognize the complicating factors in the federal system which inhibit utilization and vary from this conceptual schema. As already noted, the choice of user audience carries with it implications for the way in which the evaluation is to be conducted. Furthermore, the choice of user and appropriate use has implications for the organizational structure of evaluation services within a department. For instance, some organizations (e.g., centralized) are amenable to satisfying one kind of user need (e.g., department management), but are not at all conducive to others (e.g., program managers).

REFERENCES

Alkin, M. C. Evaluation: Who needs it? Who cares? Studies in Educational Evaluation, 1975, 1, 201-212.

Alkin, M. C., Daillak, R., & White, P. Using evaluations: Does evaluation make a difference? Beverly Hills, CA: Sage Publications, 1979.

Alkin, M. C., Kosecoff, J., Fitz-Gibbon, C., & Seligman, R. Evaluation and decision making: The Title VII experience. CSE Monograph Series in Evaluation, No. 4, 1974.

Caplan, N., Morrison, A., & Stambaugh, R. J. The use of social science knowledge in policy decisions at the national level. Ann Arbor, MI: Institute for Social Research, 1975.

Chelimsky, E. (Ed.) A symposium on use of evaluation by federal agencies, 1. Mclean, VA: Mitre Corporation, November 17-19, 1976.

Conner, R. F. The evaluation of research utilization. In M.W. Klein, & K.S. Teilmann (Eds.), The Handbook of criminal justice evaluation, Beverly Hills, CA: Sage Publications, 1979.

Florio, D. H., Behrmann, M. M., & Goltz, D. L. What do policy makers think of educational research & evaluation? Or do they? Educational Evaluation and Policy Analysis, 1979, 1(6), 61-87.

House, E. R. The conscience of educational evaluation. Teachers College Record, 1972, 73(3), 405-414.

Knorr, K. D. Policymakers' use of social science knowledge: Symbolic or instrumental? In C. H. Weiss (Ed.) Using social science research in public policy making. Lexington, MA: Heath, 1977.

Patton, M. Q., Grimes, P. S., Guthrie, K., Brennan, N. J., French, B. D., & Blyth, D. A. In search of impact: An analysis of utilization of federal health evaluation research. Minneapolis, MN: University of Minnesota, 1975.

Pelz, D. Some expanded perspectives on use of social science in public policy. In J. M. Yinger & S. J. Cutler (Eds.), Major social issues--A multidisciplinary view, New York: The Free Press, Macmillan, 1978.

Weiss, C. H. Utilization of evaluation: Toward comparative study. In C. H. Weiss (Ed.), Evaluating action programs: Readings in social action and education. Boston, MA: Allyn & Bacon, 1972.

Weiss, C. H. Knowledge creep and decision accretion. Knowledge: Creation, Diffusion, Utilization, 1980, 1(3).

CHAPTER 3

WHO CONTROLS EVALUATION? THE INTERORGANIZATIONAL COMPLEXITIES OF EVALUATION RESEARCH

Robert K. Yin
Massachusetts Institute of Technology

A. HOW EVALUATION RESEARCH IS REALLY ORGANIZED

Evaluation research teams have perpetrated a myth about themselves: that researchers alone control the quality, usefulness, and relevance of an evaluation study. The myth is reflected in the common remedies given for improving evaluation research. We are told that, if only the research was designed or conducted more carefully, the study might have been better (e.g., Berryman and Glennan, 1978). More technically, this advice is often translated into modified research designs, the search for better measures of educational performance, and the recruitment of more qualified and experienced research personnel.

This myth has been amplified by most evaluation textbooks as well as by the implicit norms of policymakers. Among typical evaluation texts (e.g., Rossi et al., 1979), the scope of coverage includes concerns about the research: its technical design and the ways of reducing threats to reliable and valid findings. Very little is said, in most texts, about the degree to which the research team may or may not control these facets of the research; the issue is rarely even addressed. Similarly, among the implicit norms of policymakers, the ways of improving the quality and utilization of evaluation studies are assumed to be matters of research techniques. Thus, for instance, the U.S. General Accounting Office (GAO) is charged with identifying improved evaluation methods--the assumption being that such methods need only be implemented by researchers in order for the state-of-the-art to improve (e.g., U.S. General Accounting Office, 1978).

In fact, the outcomes of evaluation research are not completely controlled by the research team. Instead, every evaluation study must

be regarded as a complex, interorganizational affair, involving at least three parties:

- A research team, usually located in a university or an independent research organization;
- The practitioners operating the action program being evaluated, usually located in federal, state, or local levels of government; and
- The officials sponsoring or funding the evaluation study, often synonymous with the officials funding the action program, and usually located in a federal agency.

For the purposes of further discussion, these three parties will be considered the *research team*, the *action agency*, and the *sponsoring agency*. The purpose of the following paper is to show how all three parties can be said to share the control over an evaluation study, and thus how any improvements in the quality or utilization of evaluation research will require coordinated efforts--and not just actions by the research team.

A Contrast: Traditional Academic Research

Before discussing the complexity of evaluation research, the present paper should clarify the origins of the myth. These are embedded in the traditional organization of academic research, in which a research team does indeed work independently of the other two parties.

In traditional academic research, the research team generally decides what to study and how the research should be conducted. In some cases, the research may involve an action site--e.g., a classroom, a school administration, or a governmental program--from which data will be collected. However, these action sites are selected by the researchers on the basis of the intended research design, and the participation of the action sites naturally depends on their willingness to cooperate. Often, a research team may be refused access to a particular site. But when access is granted, it is on the basis of a

mutual and voluntary agreement between the research team and the action agency. For this reason, the traditional model does rightfully focus on the primacy of the research team's role--there may be no action site, or when one exsists, the action site's participation is decided on an individual basis and is not usually part of any broader programmatic context.

Similarly, the role of the sponsoring agency in traditional academic research is minimal. The sponsoring agency, often making a grant award to the research team, takes no greater interest in the research beyond some measure of administrative accountability and research success--usually taking the form of nominal progress reports followed by formal academic publications. In this part of the relationship, the research team may actually know very little about the sponsoring agency's bureaucratic environment and procedures; knowledge of these issues is further buffered by the university within which the research team operates.

In the traditional model of doing research, then, the research team does mainly control the research. The design of the research is created and proposed by the researchers, the conduct of the study is fully under their control, and any problems with the quality or usefulness of the research can be correctly attributed to the skills of the research investigators. For this reason, textbooks aimed at improving the research design of various types of studies, or at developing better instruments and measures, are appropriate ways of improving the research.

The Inappropriateness of the Traditional Model to Evaluation Research

This very situation, in which the research team is the prime and generally only actor in the conduct of academic research, is inapplicable to evaluation research. This conclusion is based on four observations.

First, *the research team must work with a specific set of action agencies*. The designated action agencies are, of course, those involved in the program being evaluated. However, their participation in the research may not be voluntary, and whether they feel threatened by the research team or not, considerable efforts must be made--during the conduct of the evaluation study--to develop a workable relationship

between the research team and the action agency. More often than not, this workable relationship is based on a set of *quid pro quos*, of which the following are examples:

- In return for access to agency documents, the research team may have to collect certain data not necessarily relevant to the evaluation study but needed by the action agency;
- In return for using the action agency's facilities, the research team may have to use its computational facilities to produce information for the action agency;
- In return for the action agency's participation in the study and review of the results, the research team may have to assist the action agency in preparing one of *its* proposals for federal funds; and
- In return for using the time of the action agency's staff, the research team may have to provide technical assistance, of an informal nature, to the action agency.

As any evaluation researcher knows, this list can be quite long. More important, the success of the research has become increasingly dependent upon the workability of this relationship.

Second, *the sponsoring agency often plays a major role in setting the conditions for doing the research.* There are situations where research teams do initiate their own evaluation studies (e.g., see the studies reviewed by Bernstein and Freeman, 1975). However, in most large-scale evaluations in education, the studies are "procured" by the sponsoring agency (Sharp, 1980). This means that the sponsoring agency sets the major boundaries for the research, including:

- The overall level of effort to be expended in the research (note that in traditional research, this level is is determined by the research team in its original proposal);

- The scope of work;
- The types of issues, research design, and measures that are to be used; and
- The timing of various phases of the research and deadlines that are to be met.

The research team, of course, is not a completely passive actor in determining these conditions. But the increasing explicitness of the requests for proposals (RFPs) that are currently issued by sponsoring agencies means that the staff of the sponsoring agencies have been increasingly designing the "technical" aspects of the research to be done.

Third, *the sponsoring agency and the action agencies often impose limits on the research through the design of the action program.* One common occurrence is for the action sites to be selected on grounds independent of research considerations--e.g., political and administrative criteria. For instance, in federal programs, a regional distribution of action sites is often the result of a political choice; but this choice constrains the nature of the ultimate reserch design. Other decisions about the implementation of the action programs--e.g., the staggered timing for initiating work at the aciton sites--also affect the evaluation study; in this case, the research team may be unable to gather uniform "baseline" data or to conduct the research in as efficient a manner as possible. These and other characteristics of the action program, then, may all have an implicit effect on the "technical" aspects of an evaluation study, but the conditions are set by the sponsoring and action agencies, and not the research team.

Fourth, *there has been an increasing fragmentation of responsibilities within the sponsoring agency.* At least three parties, all within the sponsoring agency, may have some influence over the design and conduct of the research. These parties include:

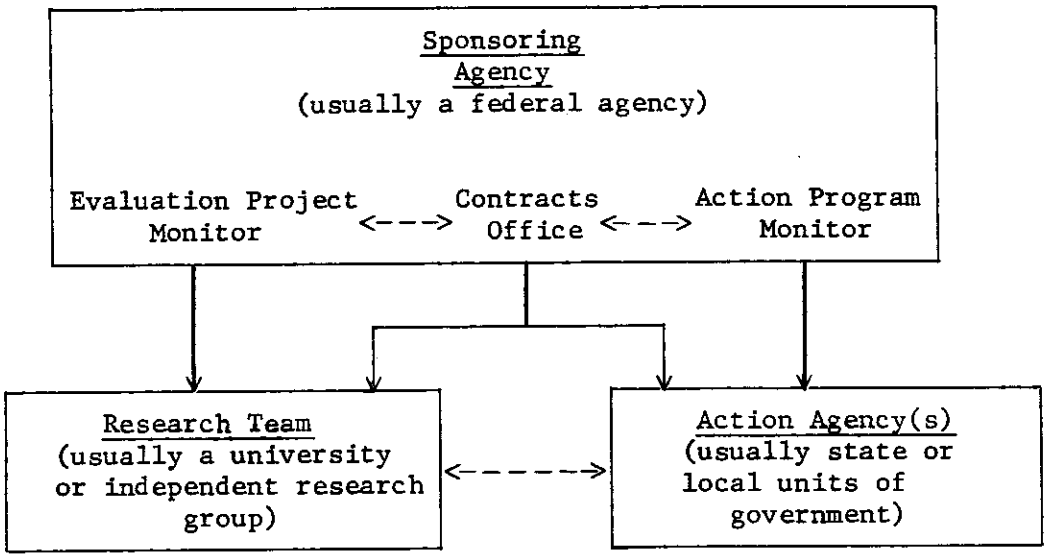
- The official "project monitor for the evaluation study itself;

- The program monitor responsible for implementing the action program; and
- The contracts office within the sponsoring agency, often dealing simultaneously with all of the other parties, within and outside of the sponsoring agency.

A research team must learn to deal with all of these parties. Sometimes, compromises must be reached because the evaluation project monitor and the action program monitor are both the audiences for the evaluation study. Other times, the contracts office can create difficulties by requiring the approval of specific activities within the action program or within the evaluation study, but then delaying action to such an extent that the research team must make further modifications in its original research plan.

In fact, this final observation regarding the fragmentation within the sponsoring agency suggests the full organizational complexity of conducting evaluation research: Three types of agencies and five relevant parties must all collaborate in order for the research to be done. These relationships are shown in Figure 3.1.

Figure 3.1
THE INTERORGANIZATIONAL COMPLEXITIES OF EVALUATION RESEARCH



General Observations

The interorganizational relationships just described, and the complexities shown in Figure 3.1, all show how evaluation research can no longer be considered along the same lines as traditional academic research. Moreover, Figure 3.1 has merely indicated the major roles that are involved in the organization of an evaluation study. In numerous specific instances, the array of actors can become even more diverse. For example, in some evaluation studies, *Congressional staff or members* may be a key part of the audience for the evaluation results. As another example, a full description of the action program might have to show the presence of *community groups*, often given an official role in monitoring the action program and the evaluation study, or the presence of a *technical assistance contractor*, whose responsibility is to help the action agencies (there may even, under certain circumstances, be a technical assistance contractor to help the sponsoring agency). Clearly, the list grows.

What all of this means is that our concerns with the outcomes of evaluation research--i.e., the quality, utilization, and relevance of the research--are not solely controlled by the research team. Textbooks and policymakers that ignore the complex interorganizational relationships are therefore inaccurate in suggesting that improvements in evaluation research can be made by the research team alone. Concerns such as "if only a better research design had been used . . ." must be considered along with other, equally relevant concerns, such as "if only the contracts office in the sponsoring agency had approved the action agency's budget in time. . ."

To illustrate the effects of this interorganizational complexity in terms of the quality, utilization, and relevance of evaluation research, the next section of this paper describes a few contemporary conditions within which evaluation research must be conducted.

B. ILLUSTRATIVE COMPLICATIONS

The interorganizational complexity that underlies evaluation research can affect evaluation studies at every major stage of their development: problem definition and evaluation design, conduct, and dissemination. The illustrative situations described below have been encountered by any number of evaluation studies in education,* and are thus described in general terms only. Each situation represents a complication imposed by the fact that several parties, and not just the research team alone, have to be involved in making the critical decisions.

Designing an Evaluation Study

One prominent complication occurs at the design stage. First the sponsoring agency and action agencies may limit the range of relevant evaluation designs through the design and implementation of the action program. Second, and more important, the basic evaluation design is described, often in great detail, in the Request for Proposals (RFP) that is issued by the sponsoring agency as the initial step in supporting an evaluation study.

Contemporary RFPs often specify the sites to be studies, the data elements to be analyzed, and the time intervals for different data collection steps. In short, the RFPs can dictate the entire scope of the evaluation design. In response, proposing research teams may attempt some modifications. However, the major modifications that will occur, if any, are likely to occur after an evaluation award has been made--when the sponsoring agency and research team are more openly able to agree about any shortcomings in the original RFP.

---

*Many specific illustrations are found in the full text of the Committee report.

The final research design of the evaluation study, then, is a function of: (a) the nature of the action program as it has been implemented, and (b) a negotiated settlement between the sponsoring agency and the research team. What this means is that if one is interested in increasing this aspect of the *quality* of evaluation research, far more issues must be addressed than the mere stipulation of methodological choices. Guidance is needed concerning such steps as:

- Conditions in the design of the action program that might negate the ability to do evaluations of minimal quality;*
- The process whereby an RFP is written and reviewed, and the staff persons who are involved in these activities; and
- The ground rules for any negotiations in creating the final design.

Of particular interest, among these steps, might be further inquiry into the training and background of the sponsoring agency's staff that is responsible for issuing RFPs. Often, one suspects that these individuals, who may have great influence over the design of an evaluation study, are either inadequately trained in evaluation or inexperienced in conducting evaluations. Often, such staff persons believe that a "textbook" version of an evaluation can be done, and they fail to recognize the actual political or administrative realities in doing the evaluation.

Getting the Evaluation Done

Similarly, the conduct of an evaluation is more complicated than the collection and analysis of data by the research team. For instance, several parties may have to review the data collection instruments proposed by the research team. The review can range from full, formal approval by the FEDAC forms clearance process to a less formal review and approval

*The author knows of no list, for instance of types of action programs that cannot be evaluated. Yet, some do exist and ought to be recognized as such.

by the sponsoring agency. In a few cases, the action agencies may also have a part in reviewing the instruments--an activity usually conducted by a "user panel" that has been established to advise the research team.

Under these conditions, the process of instrument development again becomes a negotiated process. The final product, both in scope and depth, can represent a compromise among competing priorities and cannot necessarily be regarded as the best state-of-the-art from a research point of view. Again, to improve the quality of future evaluations, guidance is not just needed on the techniques of instrument design. Systematic information is also needed, for example, on how the FEDAC process can be conducted more smoothly and in a more timely fashion--a responsibility that often involves key staff persons in the sponsoring agency. Similarly, guidance may be needed on the fair limits of non-research priorities--e.g., how far the research team should go to incorporate questions of interest to one of the other parties but not critical to the evaluation.

Reporting the Evaluation Results

The textbooks tell us that, for utilization purposes, the production of evaluation results should occur in a timely fashion. Usually, this means that the results should be made available when key policy decisions are being considered.

Not surprisingly, the research team does not have full control over this stage of the evaluation, either. Although the research team may have tried to keep to a policy-relevant schedule, the final results must also be reviewed by the sponsoring agency (and sometimes by the action agency) before a report can be made public. Delays can certainly have occurred in the conduct of the research. But the final reporting of results can also be delayed by the action of these other two agencies. Sponsoring agencies may be especially susceptible to cumbersome and lengthy review processes. For instance, the draft report may be shown to a wide variety of individuals within the sponsoring agency, all of whom may have a different point of view about the evaluation or the program being evaluated. Under such conditions, the research team often

has difficulty in merging the various comments into a coherent pattern, adding to the time and effort needed to review the final report.

Independent of the issued of whether sponsoring agencies may purposely, for political reasons, delay the issuance of final reports, there are few guidelines for a minimally acceptable review process. One of the notable gaps in most RFPs and proposals is that the sponsoring agency's review process for draft manuscripts is in fact not described at all (usually, the RFP merely stipulates that the review will occur within a particular time period). Are some review processes more desirable than others? Can the review process be streamlined so that the evaluation results can really be reported in a timely fashion? These are some of the issues that still need to be addressed and that again go beyond either the technical aspects of doing evaluations or the full control of the research team.

Summary

These few illustrations should be sufficient to indicate the degree to which evaluation research is a joint, interorganizational enterprise. As such, any attempts at improvement must not only be focused on the research team and its technical methodologies, but also on the capabilities of other relevant parties, including the staffs of the sponsoring agency and action agency.

For example, rarely has one heard any debate regarding the ways in which RFPs should be written or even by whom they should be designed. (For a modest beginning, see Weidman, 1977.) Yet, from the standpoint of improving evaluation research in education, changed in the RFP process may be more important than any potential changes in the capabilities of the research team. Perhaps it is even time for a textbook on how to write RFPs and how to monitor research, or even on the development of minimum standards regarding those staff positions in the sponsoring agencies--to appreciate better the role that such staff can have in affecting the quality, relevance, and utilization of evaluation research.

Similarly, this may now be a good time for further reviews of all the barriers confronting the conduct of evaluation research, including the role of FEDAC clearance. A few years ago, one research investigator did attempt to catalog all of the "regulations" that affect research (Gandara, 1978); the review revealed a mine field of potential barriers and problems. If evaluation research has become over-regulated, changes must be contemplated in the regulatory environment and not just in the technical aspects of research methodologies.

Finally, it is now clear that "successful"evaluation researchers are those who are able to manipulate the interorganizational complexities that have been identified within this paper. For instance, Paul Hill, who conducted a successful evaluation of the compensatory education program, wrote that the five ingredients for success included the following conditions (Hill, 1978):

1. The evaluation was aimed at decisons that "users" could make.
2. The evaluation was conducted in open consultation with potential users.
3. The evaluation recognized that research information was only one source of information that would be available to users.
4. The evaluation results allowed for divergent positions and values.
5. The results of the evaluation were produced in a timely manner, to feed debate about the action program.

What is surprising about this list of major lessons from a successful evaluation researcher is that not one of the lessons involved the technology of the evaluation. Each of the lessons, on the contrary, covers some aspect of interorganizational relationships, showing how the research team must be prepared to manage such relationships.

Myths die hard. The purpose of this paper will have been served if we no longer think of evaluation research as being organized like traditional research. Evaluation research is not done solely by a research team and

and therefore cannot be controlled by the research team alone. The staffs of other organizations, as well as the nature of interorganizational relationships, all become important to the design and conduct of an evaluation study. Recognition of this complexity should lead to better insights on how to improve evaluations.

REFERENCES

Bernstein, I. & Freeman, H. Academic and entrepreneurial research, New York: Russell Sage Foundation, 1975.

Berryman, S.E. & Glennan, T.K. Jr. An improved strategy for evaluating federal programs in education. Unpublished paper. The Rand Corporation, Santa Monica, California, 1978.

Cronbach, L.J. et al., Evaluation for an open society, Forthcoming.

Gandara, A. Major federal regualtions governing social science research, The Rand Corporation, Santa Monica, California, March 1978.

Hill, P.T. Evaluating education programs for federal policymakers: Lessons from the NIE Compensatory Education Program. Unpublished paper. The Rand Corporation, Santa Monica, California, December 1978.

Rossi, P.H. Researchers, scholars, and policy makers: The politics of large scale research. Daedalus, January 1964, pp. 1142-1161.

Rossi, P.H. et al., Evaluation: A systematic approach. Beverly Hills: Sage Publication, 1979.

Sharp, L.M. Performers of federally-funded evaluation studies. Paper prepared for Committee on Evaluation in Education, National Academy of Sciences, Washington, D.C., July 1980.

U.S. General Accounting Office. Federal program evaluation, Washington, D.C., October 1978, PAD-78-83.

Weidman, D.R. Writing a better RFP: Ten hints for obtaining more successful evaluation studies. Public Administration Review, 1977, 37, 714-717.

CHAPTER 4

EVIDENTIAL PROBLEMS IN SOCIAL PROGRAM EVALUATION AND THEIR ANALOGS FROM ENGINEERING AND THE PHYSICAL SCIENCES, MEDICINE, AND AN ASSORTMENT OF OTHER DISCIPLINES, BASIC AND APPLIED, TOGETHER WITH SELECTED HISTORICAL REFERENCES[1]

Robert F. Boruch
Northwestern University, Evanston, Illinois

1. Introduction

This paper concerns the problems that are engendered by efforts to collect evidence about a problem or a proposed solution. The special focus is on problems common to both social program evaluation and to evaluation in other arenas, notably the physical sciences, engineering, medicine, and occasionally commerce. Not all of the problems are new, despite contemporary arguments over what to do about them. And so the paper is studded with references to early pertinent work.

I have several reasons for developing a disquisition of this sort. In the first place, science in the abstract recognizes that the difficulties of accumulating decent information transcend discipline. But the lay public and its representatives and analysts in sundry disciplinary camps often do not. The failure to apprehend that the same problems occur in both physical and social sciences, indeed that many are very durable, is a bit shameful. It results in social research's being construed as more feeble than work in other vineyards. It is more feeble in some respects. It is at least as robust in others though. And crude comparative work of this sort may help to illustrate the point.

The more immediate, less rhetorical, feature of the motive concerns the responsibility of agencies, such as the U.S. General Accounting Office, to oversee performance of government in a variety of tasks. The problems that these agencies encounter are often statistical and scientific at their core, though they are infrequently labelled as such, and common to several disciplines. This paper may help to remedy this problem as well.

A second motive is more personal. The writer is a metallurgical engineer turned social scientist. The vernacular differences I encountered in stumbling from one field to the other are tedious at best. The social program evaluator's "formative evaluation" is no different, though perhaps more obscure for good or ill, from the engineer's "trouble-shooting" or "development." At worst, they often imply notional differences, between qualitative and quantitative, subjective and objective, that are often gratuitous even obstructive. In this respect, the spirit of the paper is akin to others, notably Florman's Existential Pleasures of Engineering. It is more a catalog than intellectual history or dialectic between camps. But if it succeeds in stimulating better understanding of the nature of such problems, one of its objectives will have been met.

The examples illustrate failures of scientific knowing or of common sense, little cortical collapses. They are not intended to demean research in the physical sciences, medicine, or business. The point is that the problems are persistent, and we ought to appreciate their appearance in a variety of human enterprise.

2. Implementing Programs and Characterizing Delivery

It is something of a truism that social programs are never delivered as advertised. The social scientist often finds it very difficult to assure that the program under investigation has the form that it is supposed to have. Moreover, it is often difficult to monitor the discrepancy between plan and its actualization systematically. The problem is persistent in evaluating complex, broad aim efforts, such as Model Cities Programs during the 1960's (Pressman and Wildavsky, 1973). It is characteristic of newer evaluations, including those directed at programs which are said to be well structured but structure depends heavily on individuals' following marching orders. They often do not or cannot. See, for instance, Fairweather and Tornatsky (1977) on evaluations in mental health, Kelling (1976) on police research, Sechrest and Redner (1978) on estimating the effects of innovative criminal rehabilitation programs, and Rossi (1980) on education, welfare, and other programs.

Laboratory research is not spared the problem, of course, though its severity and measurability differ from the field variety. Grey eminence

L. L. Thurstone encountered military trainers who gave instruction secretly to control group telegraphers in the interest of assuring they got the benefit of sleep learning (Mosteller, 1978). Partly on account of such early difficulties, it is common practice in social psychological research to check treatment manipulations. The measurement and reporting problems apply to methodological research on improving cooperation in mail surveys (see Good, 1978, on Christopher Scott), to educational evaluation (Leonard and Lowry, 1979), and to applied social research on other topics (Boruch and Gomez, 1979).

The problem is not confined to evaluation of social programs. It appears in the engineering sciences where, for example, allegations that reliability of control over variables affecting reactor cooling systems have been a grave concern (Primack and von Hipple, 1974). The control problem in some chemical processes has been sufficient to warrant V.V. Federov's developing new approaches to understanding in randomized tests at Moscow. (Despite this there appears to be little attention to the problem in texts of experimental design in industry.) Bureaucracies have simply forgotten to implement plans for pesticide control (U.S. General Accounting Office, 1968) and to deliver vasectomy kits in fertility control programs (Sullivan, 1976). They have denied the existence of treatments or mislabelled them: Recall the U.S. Defense Department's denial of the use of poison gases at the Dugway facility. The problem is implicit in early agricultural experimentation as well if we judge correctly from Yates' (1952) concerns about correction of bias in moving from laboratory versions of fertilizer application to field studies. It is also buried in the history of manufacture, including the production and adulteration of foodstuffs: recall Accum's treatise for the 19th century consumer. Lest the blame be laid on institutions, recall that the odds on being treated by the pill advertised on the label are 9 to 1 according to the Food and Drug Administration. What happened to the remainder is not known.

There are more than a few interesting parallels between evaluation of social programs and meteorological studies of the past ten years, judging from Braham (1979), Kruskal (1979), Flueck (1979), Crow, et al. (1977), Neyman (1977), and others. The commonalities are especially evident from randomized tests of the effects of cloud seeding on precipitation. Pilots who were responsible for seeding silver iodide crystals had their own

preferences about where to fly, notably in sight of the coastline for the Israeli experiments. Decisions had to be made about whether to shift the target accordingly. Spillover of seeding or contamination of neighboring clouds is a threat to the validity of inferences in these studies, just as it is in the social sector where children not assigned to special education may receive it anyway from well-intentioned teachers. Seeding flares in early experiments in Florida were imperfect just as nutritional supplements were in the early Colombian experiments on the supplement's effect on ability. Measuring the level of imposition or of receipt of treatment seems to be no less difficult here than in the social sector. Indicators of intensity of treatment, for instance, are sometimes crude, e.g., recorded duration of seeding and mean wind speeds in the target area. Reliably indexing cloud conditions is all but impossible on account of their variability, and this problem is analogous to the chronic one of assaying the local conditions that may affect delivery of welfare services, educational TV, or income transfer payments, in evaluating social programs.

The unwillingness or inability of field staff to adhere to regimen demanded by a new social program seems not much different from the reluctance evident in some tests of medical innovation. For instance, attempts to determine whether conventional, enriched-oxygen environments for treatment of premature infants actually caused blindness met with remarkable resistance from some nurses and physicians. The latter were unable to countenance depriving infants of oxygen, though subsequent research demonstrated that oxygen was indeed influential in producing blindness (Silverman, 1977). The difficulty here parallels earlier ones, encountered by British Army Surgeon General John Pringle and others who attempted to reform the sanitation practice of hospitals (Marks & Beatty, 1976). The problem also extends to well-trained specialists where, for example, the integrity of an operation such as coronary bypass is variable judging by indices such as perioperative heart attacks, graft patency, and crude hospital mortality rates (Proudfit, 1978). A similar problem, in less obvious form, emerges when one considers the material used in tests of vaccines and drugs. Confirmatory tests of polio vaccine were disrupted briefly by a product that induced poliomyelitis instead of preventing it (Meier, 1972; Meier, 1975). The Indian tuberculosis prevention trials

were executed partly to determine whether effectiveness of vaccine, demonstrated earlier to have been effective, had altered because of strain mutation and changes in antigenicity, variation in production methods or in doseage levels (Tuberculosis Prevention Trial, 1979; Altman, 1980).

A cruder form of the problem involves receipt of treatment and adherence to regimen. For example, in the Kaiser-Permanente tests of multiphasic screening, many of the individuals assigned to the screening program failed to turn up for periodic examination. The research staff, interested in effectiveness of screening and not of natural turn out rates, mounted an intensive telephone program to encourage participation in the free and presumably beneficial service (Cutler et al., 1973). Similar encouragement strategies have been necessary to obtain interpretable estimates of the effects of viewing educational television. A good deal of the argument over the implications of the University Group Diabetes Program hinges on an identical problem--a minority of patients in at least one group appear to have adhered faithfully to the treatment regimen to which they were assigned (Kolata, 1979b).

3. The Odds on Success and Failure and Uniformed Opinion

> These were the generations of Budgeting
> . . . . . .
> Planning-Programming-Budgeting begat Management by Objectives
> Management by Objectives begat Zero base Budgeting
> Zero base Budgeting begat Evaluation
> Evaluation begat Experimentation
> Experimentation showed that nothing works.
>
> From A. Schick, Deuteronomy.
> The Bureaucrat, 1976.

The concern that innovative social progams will fail is justified. But the expression of that concern is often pessimistic, occasionally alarmist in some camps, wildly optimistic in others. At George Washington University, for instance, we were taken aback by the plaint that evaluation is discouraging to the public, bureaucrats, and politicians because positive effects appear infrequently, and so it should be trimmed. In one of Patricia Graham's public addresses as head of the National Institute of

Education, she took pains to recognize declining agency morale and attributed it partly to the conduct and results of contemporary program evaluations. Pessimism is not an uncommon theme in the academic sector either. Here, it is easy to find comfortable cynicism about the lack of good evidence, and occasionally, the judgment that because evidence is poor in quality, programs are also poor in quality.

One problem, of course, is to determine when the pessimism is warranted. We believe it is generally misleading as the incontinent optimism of early programs. In particular, the vague negative view is not well justified simply because we do not yet have reliable information on the relative frequency of failure, success, or mixed results of new projects. The short history of evaluative policy and briefer development of competent field testing account partly for the scarcity of data about odds. To illustrate one approach to understanding in this context, consider Gilbert, McPeek, and Mosteller's (1977) examination of high-quality evaluations of surgical innovation. Considering only well designed evaluations, they find that about one-third of such innovations are fairly successful relative to standard surgery, a third are worse than standard, and a third do not differ appreciably in effectiveness relative to normal practice. As one might expect, similar problems have affected the introduction of new drugs though current success rate is not clear. For instance, a massive reevaluation of the efficacy of drugs was undertaken by the National Academy of Science following the 1962 Drug Amendments Act. The report suggests that about 7% of the drugs and 19% of the claims were ineffective (see Hutt's remarks, page 228, in National Academy of Sciences, 1974). The most pessimistic estimate includes drugs that are only "possibly effective" and drives this statistic up to 60%. Gordon and Morse's (1975) coarser review of the well-designed evaluations which have been reported in the sociological literature suggests that 75% of the programs under study fail to detect any improvement over comparison programs. If one admits poorly designed evaluations in the calculations, the odds change of course. Some examples are given in the section on inept design of evaluations.

No comparable efforts to assay likelihood of success have been completed in education research and development. But a crude upper bound

might be obtained from statistics on projects that have passed muster with the Department of Education's Joint Dissemination and Review Panel (JDRP). The JDRP reviews evaluative evidence on projects submitted by project managers to determine if evidence and size of the project's intended effect are sufficient to warrant further federal support. It is a biased sample of all such projects since submission to review is voluntary. About 60% have been approved in recent years. At least one lower bound estimate for one category of projects is implied by a recent American Institute of Research review of bilingual programs. Only 8 out of 175 were judged to have sufficient evidentiary support to warrant approval (Boruch & Cordray, 1980).

Judgments about failure rate in the social realm are often based on what appears to be the absence of failure or mixed results in others. So, for example, the critic may point to innovations in engineering as a remarkable standard against which social innovation do not fare well. That standard is misleading in several respects not the least being general ignorance of failure rate. Ordinary bridges, for instance, do collapse. It was not until 1636 that the first quantitative treatment of stress in bridge structure appeared, written by Galileo. "Before his time the strengths and deformations of structures were determined primarily by trial and error. A structure was built. If it stood up, well and good. If not, then the next structure was made stronger where the first one failed, and so on" (Borg, 1962, p. 4). They failed at a rate of 25 per year following the Civil War. In the 1900's, bridges large enough to symbolize a new industrial age collapsed before completion because "large steel members under compression behaved differently than the smaller members that had been tested time and time again" (Florman, 1976, p. 32). Suspension bridges have stumbled since 1741 despite their stately grace. The failures recorded in 18th century Scotland continued in 19th century England and in 20th century United States.[2] The Tacoma Narrows Bridge, which failed in 1940 on account of progressively amplified wind vibrations, is a common illustration in introductory physics texts. The spirit of that illustration also underlies examples of flaws in the evaluation of Headstart, cited in graduate texts on design of evaluations. The rules for making bridges robust against amplified vibration did not become clear until the 1950's.

Resnikoff and Wells (1973) catalog examples and explicate the rules in their delightful mathematic text. The new arena for failure here appears to be bridges remarkable for their length or age. Up until a few years ago, for instance, the Tampa Bay's Sunshine Skyway had "ample clearance for even the largest ocean-going vessel." The fact that the tallest supports did not collapse when many of the rest did is curious but no help at all to anyone who wants to cross the bay by auto. Winds up to 100 mph swept away a sizeable chunk of the new Hood Canal pontoon bridge in Washington state, and a size fraction of the $30 million investment with it (Los Angeles Times, February 14, 1979, p. 1, 8). Maintenance failures and deterioration may hasten the demise of New York's Queensborough (59th Street), the Golden Gate, and others used as bad examples in Congressional testimony on the 1978 Highway Act.

If we examine the start up of businesses, we find prospects for failure no less formidable. For 1978, the ratio of business failures to start-ups was about 20%. This estimate is a conservative one since Dun & Bradstreet, the repository for such information, defines business failure as a voluntary action involving loss to creditors or court proceedings--bankruptcy. D & B maintains "that every year several hundred thousand firms are started and almost an equal number are discontinued" (Dun & Bradstreet, The Business Record Failure, 1979, p. 3). But commercial enterprise is certainly better off now than during the late 19th century. The commercial death rate, as it was labelled at the time, was double the number of new businesses added. The ratio of failures, defined in terms of liability, to start-ups ranged from 15% to 90%, or so said Bradstreets's (Stevens, 1890/1891).

The point is that, contrary to the opinion one may develop based on anecdotal reports, narrow personal experience, or poorly designed evaluations, innovations in a variety of areas succeed less than half the time, and probably a good deal less than a third succeed at the field test stage. Innovative educational programs may succeed at roughly similar rates when properly evaluated.

Despite the occasional appearance of big bang effects, advances in any science are usually small. This makes designing evaluations which are sensitive to small effects very important. Given a design which provides

some protection against competing explanations, anticipating the likelihood that program effects will be detected, if they occur at all, is reasonable. But it is difficult to find formal power analyses in educational evaluations, making it difficult to determine if the design was indeed sensitive. It is small comfort that the same problem, ignoring a fundamental technology affects medical research (Freiman, Chalmers, Smith, and Kuebler, 1978) and less recent research in psychology (Cohen, 1962). That the technology, even where occasionally exploited, is often based on optimistic rather than realistic guesses about program effect size is even less comforting (see Daniel, 1972, for instance, on industrial experimentation).

4. Reliability and Validity of the Data

Mark Twain, according to Mark Twain, was not terribly bright. But he did have the wit to assay reliability and validity of phrenologists' readings of his skull and palmists' readings of his paw. Some readings were wildly unreliable: Bumps interpreted one month disappeared entirely or became dents on the second engagement. The most reliable palmist appears to have averred repeatedly that Twain had no sense of humor (Clemens, 1917/1959).[3] Some researchers exhibit a sturdier indifference to common sense.

Projects without much concern for quality of information include the Philadelphia Federal Reserve Bank's evaluation of the Philadelphia School District, the Federal Aviation Administration's evaluation of the Concorde's impact on communities in the airport's vicinity, and many of the recent studies of the impact of desegregation. For these and other cases, establishing the quality of a response measure is essential for obtaining a decent description of the nature of a social problem and for estimates of the effects of a program on the problem.

Especially when evaluations are used to inform policy, the consequences of ignoring flaws in the information can be serious. In covariance analysis of observational data, for instance, simple random errors of measurement can bias estimates of program effect. Under conditions commonly found in the field, the result is to make weak programs look harmful in compensatory education (Campbell & Boruch, 1975) and manpower training (Borus, 1979), and to adulterate evidence about sex or race discrimination

in court cases. Similarly erroneous conclusions may be drawn in applications of the same method to basic research data on schizophrenia, for instance (Woodward & Goldstein, 1977). The difficulties abide for anthropological disciplines as well as their more numerical sisters. Recall for instance Lienhart's view that Darwin was misled into believeing Terra del Fuego natives were cannibalistic by natives who wished to be entertaining and cordial. (See Przeworski and Teune, 1970, for illustrations and a bibliography.)

It is not difficult to find analogs to simple problems of reliability of measurement in medical diagnoses. During the 1960's, for instance, well-informed physicians knew that simple tests for gonnorhea yielded false positives. One physician, not so well informed, managed to start an outbreak of mass psychogenic illness (contagious hysteria) among high school students by simply failing to read medical literature. Understanding the traps in simple tests led Mausner and Gezon (1967) to avoid relying on vaginal smears alone and ultimately to their development of a remarkable case study of the episode. Measurement error in the response variable appears now in more complicated ways, judging from the University Group Diabetes Program. There, not a little of the ambiguity in evidence is attributable to the way diagnosis of cardiovascular disease depends on whether one conducts an autopsy. And, of course, the random instability in blood pressure, among other traits, causes no end of argument about who is hypertensive and who is not, and about whether labile hypertension is indexed by blood pressure is merely regression to the mean or similar artifact of the way we measure or respond to measurement over time (Kolata, 1979). The problem is a hoary one in medicine and well-documented at least for illnesses such as smallpox and measles. Still, it is a bit unnerving to stumble over examples: Citizen Graunt inveighed against the "ignorant and careless searchers" who did not accurately enumerate deaths in the 1600's. His little catalog of ways that cause of death might be misconstrued (does a seventy-five year old man die of "the cough" or of old age?) is a rudimentary theory of misclassification (Graunt 1662/1973).

Just as judgment about children may be influenced by teacher's expectations, medical assessments are sometimes slanted by physician's expectations more than by evidence. Recall that in McEvedy and Bear's

(1973) study of neuromyasthenia, symptoms similar to those exhibited by victims of poliomyelitis were also exhibited by physicians and nurses without the latter disease. The physicians regarded neuromyasthenia as a clinical syndrome when indeed the problem was psychogenic. Barnes' (1977) fascinating review of worthless surgery is also pertinent here. He reminds us that ptosis was characterized early in the 20th century as a condition in which the position of internal organs was "abnormal." Surgeons thought the abnormality caused a wide variety of symptoms. Kidney displacement, for instance, was alleged to produce neuroticism, back pain, and vomiting. We know now that ptosis is not an organic problem, that surgery was unwarranted, and that diagnosis and etiology were nonsense. The reader may think this illustration far fetched. It is not, judging from recent efforts to slice the incidence of tonsillectomies, hysterectomies, and adenoidectomies (see Dyck et al., 1977, for instance).

The engineer has to accomodate problems of error in measurement too, of course. And despite the awesome growth of the instrumentation industry, they are often no less severe. For instance, in the Handbook of Dangerous Properties of Industrial Materials, Herrick (1979) reports that the reliability of air screening is such that readings are within ±25% accuracy. This suggests that reports of environmental tests should routinely provide information about their reliability, just as one ought to provide estimates of reliability of personality inventories, questionnaires, and the like. The validity of environmental test results depends no doubt on local circumstances. And it's conceivable that the results ought to be adjusted for these just as are standard measures that are influenced by temperature and buoyancy in the case of weight. The difficulty of adjusting for temperature expansion can be traced to Michaelson's efforts to correct for thermal expansion in estimating the speed of light and his failure to correct for temperature influences on the index of light refraction (Eisenhart, 1968).

Flaws in observation and measurement on a much larger scale are not advertised much, especially if they concern the military. But remarkable ones surface occasionally. Detecting an atomic blast, for example, is not as easy nor reliable a process as one might expect. The Vela surveillance satellite "saw" an explosion in 1980. What was thought until then to be a unique signal, associated with a blast, turns out not to be

unique. It is produced by peculiar confluence of natural phenomena as well. The influences on accuracy of measurement are as difficult to assay in military engineering. Part of the controversy over the Airborne Warning and Control System hinged on the airplane's suceptibility to attack and the system's lack of robustness against state-of-the-art devices for signal jamming (Sousa, 1979). And, of course, the interest and skill of individuals given responsibility for measurement plays a major role. The federal delegation of authority to state governments in the national dam safety program, for instance, resulted in data which varied enormously in quality. Dams were missed entirely in their inventory, hazards ignored, and data was inaccurate in other respects (Perry, 1979).

Recognition of such problems in the sciences is not recent. Galileo had the sense to have the ball descend the channel repeatedly to assure that his estimates of acceleration rate were decent. Not more than 30 years later, Graunt (1662/1973) issued complaints about the indifferent quality of records available for political arithmetic. Over a hundred years later, astronomer Simpson made the same point in writing about the need to obtain a mean in observations. Echoes of that advice can be detected in at least one electrical engineering text of 1917 and one chemical engineering text of 1938. (See Eisenhart, 1968, for a remarkable treatment of the topic and for references to these examples.) As one might expect, there are physical antecedents to contemporary debates over difinitions of intelligence, ability, and the like. The difference between the American inch and the British inch, created by legal fiat in 1866, was small but caused no end of problems until 1966 when both were defined by agreement as 2.54 cm (Barry, 1978).

The idea that there are important qualitative aspects to the problem of measurement error is not especially new. A founding father of statistical quality control methods recognized it in the 1930's, stressing that people, the physical devices, and other influences on measurement need to be recognized. His observations were presaged by astronomer George Biddel Airy in 1861 who warned against "light" assumptions about presence or absence of constant error, and recognition of chance variation. The statistician Gosset (aka Student) recognized higher consistency among measures taken within a day relative to those across days, and speculated

on the reasons for the phenomena in 1971 (Eisenhart, 1968). Mosteller (1978) notices similar structure in time lapse data generated during the 1860's under the support of the U.S. Coast and Geodetic Survey. Working on Gosset's turf in 1956, Cunliffe (1976) found notable random variation and peculiar within-laboratory variation in measures of the volume of Guiness beer in bottles. This was apparently remarkable enough to justify "very delicate conversation" between Cunliffe and Guiness's chemist, from which each "retired, somewhat wounded."

The little herd of theories and inventions which helped to improve understanding of the qualitative aspects of measurement in physical and engineering sciences seems not to have been matched in the social sector. But some relevant work has been done. In broadening his thesis on social experimentation, for example, Campbell (1975) espoused a side theory on corruption of social indicators. The idea is that as soon as it becomes well known that a measure is being used in making policy decisions, notably in program evaluations, the measure will be corrupted in some degree. A related idea characterizes 14th century India's Ibn Kaldun's observations on his predecessor's exaggeration of numbers in description. Numeric sensationalism exalted the status of historian and statesman then as it does now, and Kaldun's attributing the problem to lack of conscientious criticism seems no less pertinent now. During the same period, China regarded the problem of suppression of facts in censuses as serious enough to justify beheading minor officials (Jaffe, 1947). To get much beyond the idea, one must identify the main influences on corruption. For Knightly (1975), in what must stand as a model of crude theory in war reporting, this meant tracing the quality of battle statistics, from the Crimean wars to Viet Nam, as a function of incompetent journalists, self-interested generals, self-serving politicians, and as a function of what he regards as a minority, the virtuous members of each camp. Sound misreporting in recent wars seems not to have impeded military careers of some generals (Halberstam, 1969).[4]

The scholars' observations on corruption are clever and important. But it does seem sensible to recognize other persistent sources of distortion. Indifference and inability may not be as titillating as corruption but they are likely to account for more of the problem. The indifference

was recognized by Graunt if we interpret correctly his concerns about London's ignorant and careless searchers. They are implicit in Barnas Sear's reservations, as Secretary of Education for Massachusetts, about the quality of educational statistics, in 1850: "Those who know the summary manner in which committees often arrived at their conclusions in respect to this (numbers of children in various types of schools), will use some degree of caution in reasoning from such data" (Kaestle & Vinovsky, 1980). Inability is harder to infer. But it's not an implausible reason for distortion in Chinese censuses of the 14th century and afterwards: the individual being counted might regard the act as depleting one's spirit, it's something of an embarrassment to have an unmarried, marriageable daughter in the household, and so on (Jaffe, 1947). And it accounts, at least partly, for poor statistics on some diseases: 17th century attitudes toward venereal disease and its recognition appear to have been almost phobic, and probably helped to enrich the physicians of the period.

5. Access to Data and Reanalysis

Routine reanalysis of data from program evaluations is a relatively new phenomena. But the general notion of secondary analysis of social statistics is not. In the United States at least, it was implicit in Madison's arguments with Congress about the multiple uses of census information (Cassedy, 1969). It was dramatically explicit in arguments over social statistics just before the Civil War. Congressional criticism of printing contracts for the 1840 census results and John Quincy Adam's interest in census inaccuracies led to the American Statistical Association's investigating the data (Davis, 1972; Regan, 1978). There was considerable controversy since the statistics were used by slavery advocates such as John Calhoun to support the "peculiar" institution. The spirit of the enterprise in the laboratory has been durable. It is reflected, for instance, in reanalysis published in 1929, of psychophysical data generated in 1873 by C.S. Pierce. Mosteller (1978), who provides the references, rummages still further, in the interest of illustrating the character of nonsampling error.

In recent years, good illustrations stemmed from evaluative research on social programs. This includes fascinating reanalyses of Coleman's Equality of Educational Opportunity Surveys, appearing in a volume edited by, of all things, a senator and a statistician (Moynihan & Mosteller, 1972), of data from evaluations of Sesame street (Cook et al., 1975), Head Start (Magidson, 1977), and others. At times, the results are both surprising and important. Leimer and Lesnoy (1980), for instance, appear to have discovered a fundamental error in the 1974 work by Martin Feldstein, current president of the National Bureau of Economic Research. The original work, used as a basis for policy, purported to show that social security had a large negative effect on individuals' savings. The reanalyses show no such effect and imply remarkably different policy. In Fowler vs North, the Supreme Court used an economist's estimates of the effect of capital punishment on homicide rate in reaching its decision on constitutionality of that punishment. At least one major reanalysis, done after the decision, suggests that contrary to earlier conclusions, capital punishment does not have a substantial deterrent effect (Bowers & Pierce, 1980).

Similarly remarkable changes in views come about occasionally in reanalysis of physical data. For instance, a health physics laboratory recently analyzed rainfall samples following a suspected nuclear explosion. Their findings on water borne fission products appeared to confirm the fact that the blast occurred, but independent tests suggested no such thing (Science, 1980, 207, 504). The original finding appears to have been due to contaminated instruments. In the case of the University Group Diabetes Program, a federal decision to require warnings on the use of tolbutamide was made before the research was reported in professional forums and much before data was to have been released for secondary analysis (Kolata, 1979b). The requirement was eventially rescinded when arguments over implications of the data became serious.

The reanalysis of evaluative data carries no guarantee that it will inform any more than does reanalysis of other kinds of data does. Nor will it always be apparent that reanalysis will be more informative than primary analysis. Ingenuous optimism about the latter appeared among some turn of the century professors and I see no reason to ignore that history

and its implication. In particular, it is an unwarranted expectation that "as a multiplication table should be reliable for both the Tory and the Communist, the conclusion of social trends should be valid alike for the radical and conservative" (Odum, quoted in Barnes, 1979, p. 62). The data will, for example, be used for purposes other than those for which it was collected, properly and improperly. Chambers (1965), for instance, recounts how the correspondence between time series data on small pox incidence and on vaccination campaigns were interpreted by antivaccinationists as a demonstration of the invidious effect of vaccination when in fact, the campaigns were mounted following the onset of an epidemic. Barnes (1979) reminds us that Marx used data from Her Majesty's inspection of factories in ways "undreamed of" by the government. Debate about what the data mean can be extended. The UGDP trials ended in 1968, but papers which purport to find the vitiating flaw in original interpretation continue to appear (Kilo, Miller, Williamson, 1980). Fifteen years after randomized field tests of cloud seeding in the United States, arguments about what the conclusions ought to be persist (Braham, 1979; Neyman, 1979). Durable debates are not less easy to find in educatinal program evaluation though they seem to be less grim and certainly less vituperative than those in the medical arena. Magidson (1977) builds more plausible models for estimating that program's effect in 1966 or so. The models seem not to have satisfied other scholars publishing in Evaluation Quarterly since then.

It has not always been easy to secure data for secondary analysis in any of the sciences. Proprietary interests, declared or not, seem to account for data not being manifestly available to independent analysts when the North decision on capital punishment was reached. Indeed, the first major criticism of the analyses used in the case was based on conscientious reconstruction of the data from disparate sources checked to assure that the data were similar to if not identical to the information used in original analyses (see Bowers and Pierce, 1980). The territorial imperative and personal differences among scientists appears in over half the chapters of Watson's Double Helix as obstacles to fitting better models of DNA structure to raw data, X-ray diffraction photographs. The problem of access in the natural sciences is sufficiently tantilizing to warrant the attention of independent authors as well. Koestler's (1971) description of the tangle

over Kammerer's research emphasizes the difficulty of acquiring the toads he used or parts thereof to verify the scientist's claims.[5]

When the data are held by an institution, matters become very difficult indeed and may involve the courts. The problem is less one of discipline difference than contest between the government staffer and civilian. "Sharing information does not come naturally to the policy maker because knowledge is power" or so sayeth Yarmolinsky (1976, p. 265). Threats of legal suits under the Freedom of Information Act have been used to extract social data from ADAMHA, just as they have been used by physical scientists to obtain information from the Atomic Energy Commission on licensing criteria and from the Federal Aviation Administration on the Supersonic Transport (Primack & von Hipple, 1974). The Department of Defense's refusal to disclose actual sites of herbicide spray in Viet Nam impeded the attempts of the American Association for the Advancement of Science to verify the Department's claims that effects of spraying are negligible and to assess the laboratory data on the topic (Primack & von Hipple, 1974). In Forsham vs Harris the access issue commanded the attention of the Supreme Court. There, the suit brought by independent analysts argued that data generated in the University Group Diabetes Program trials should be made available for reanalysis. A period of groping among federal agencies to determine which one had the data was followed by a legal suit. Apart from the general scientific justification for access, it was argued that the data from a publicly supported project were used as a basis for major policy decision and this implied that the information ought to be made available for reanalysis. The Court ruled against forced disclosure.

Institutional reluctance to disclose information is not new of course. But it may come as a surprise that paragons of early statistical virtue, such as John Graunt, were not disposed to free access. In his introduction to Natural and Political Observations (1662/1973), Graunt advocated England's keeping records universally on burials, christenings, and an assortment of other events. But he adds ". . . why the same (statistics) should be made known to the people, otherwise then to please their curiosity, I see not" (p. 12). At the end of the monograph, he passes the buck: "But whether the knowledge be necessary to many or fit for others, then the Sovereign, and his chief Ministers, I leave to consideration," (p. 74) presumably of these

same authorities. Graunt's unwillingness, or at least ambivalence, to disclose information was not unusual. De Santilla (1955) reminds us of the "Phythaogeanprivacy of research" that characterized views of Copernicus and Galileo. Neither they nor their contemporaries were much inclined to publicize some of their observations and the constraints of religion seems to have been only part of the problem. Lecuyer and Oberschall's (1968) fascinating review of the history of social research in western Europe suggests swings between openness implied by government ordinances requiring registry publication of births, deaths, and so on in the 17th century, and the secrecy implied by surveys and reporting systems for taxation and military conscription of the 18th century. Nor does this seem to be a European phenomena. The secrecy that characterized storage of demographic data collected in 17th century Dahomey and in China in apparently all censuses is military in its origins. For Dahomey, this was probably less easy to do than it sounds: counts were represented by large sacks of pebbles and updated often.

In social statistics generally, there have been recent efforts to make information more readily available. Flaherty (1978), for instance, took a leadership role in getting international agreement on principles of disclosure, principles which run counter to conservative tradition of statistical bureaus in Britain and Germany among others. In the United States, there have been more than a few very recent efforts to assure that evaluation data are more readily available for review. Federal, rather than state, agencies, in criminal justice research, education, and census, have developed policy and are testing it (Boruch, Wortman, Cordray, 1980). The same spirit is evident in recent advice to medical researchers that the need for secondary analysis of experimental data be stored for use by independent analysts (Mosteller, Gilbert, & McPeek, 1980), the creation of data repositories for studies in meteorology (Braham, 1979), energy, environment, and others (Kruzas & Sullivan, 1978).

6. Individual Privacy and Confidentiality[6]

History

Despite contemporary rhetoric, the privacy questions that emerge in social research efforts are not new. We can trace public concern about census surveys to 1500 B.C., when in Exodus (30:11-17) and Samuel (2 Sam. 24:1-5), we find both God and man opposing military demography. Popular objections are rooted at least as much in a wish for intellective privacy as in a desire for physical self-preservation, and they are no less evident in the early history of social research in the United States. An interest in sustaining at least some anonymity with respect to the government reveals itself in colonial New England's restricting the collection of data for "public arithmetic" to publicly accessible information (see Cassedy 1969 and Flaherty 1972). The privacy theme is implicit in Madison's arguments with Congress over what data should be collected in national censuses for the sake of managing the republic. It is explicit in Lemuel Shattuck's reports to the Massachusetts Sanitary Commission in 1879, which refer to public concern about the propriety of the then-novel epidemiological survey and about the government's use of the resulting data. Shattuck's work foreshadowed controversy over routine surveys of public health and the creation of archives containing information on mortality and health during the late 1800s (Duffy, 1974). That controversy is no less apparent today in some developing countries, where for example, deaths may go unreported on account of privacy custom, memory lapse, or inheritance taxes. The collection of economic data has run a similarly difficult course, with public demonstration against the Social Security Administration's record keeping during the 1930's reflecting a concern not only about personal privacy but, from commercial quarters, also about institutional privacy.

That data obtained for statistical research ought to be maintained as confidential is probably at least as old an idea. But aside from the fine work of Flaherty (1972) and Davis (1971, 1972), there is scant historical documentation on the matter. In America at least, the idea is explicit in guidelines issued in 1840 by the Census Bureau, requiring that census enumerators regard as confidential information obtained from their respondents (Eckler, 1972). Indeed, the history of attempts to make

certain that the respondent's fear of disclosure would not inhibit cooperation in social research can be traced throughout much of the U.S. Census Bureau's existence. As the amount of information elicited grew from the simple enumeration of 1790 to the economic and social censues of the early 1900's, and as the quality of surveys shifted from the astonishingly inept efforts before 1840 to the remarkably high-caliber work of the present day, so too did the laws governing disclosure--from rules demanding public posting of information elicited in a census to explicit statutory requirements that information on individuals remain completely confidential in the interest of preserving the quality of data available to the nation. The same theme is evident in the early development of economic welfare statistics, notably under the Social Security Administration. The problem of deductive disclosure is not a new one either. Ross Eckler's (1972) history suggests that the risks of accidental disclosure based on published statistical tables, most evident in the census of manufacturers, were officially recognized as early as 1910.

Legislative protection has, in the case of the census, been helpful in resisting pressures brought to bear on this public interest by other public interests. The U.S. Census Bureau has successfully staved off demands for information on identified respondents that range from the trivial to the ignominious. The latter include attempts to appropriate census records during World War II in an effort to speed up Japanese internment. There have been requests that were superficially worthy, including location of lost relatives, and others that were not so worthy. But the same level of protection in one quarter may serve as a barrier in another. Under current rules, one may not access census records that are under seventy-two years old for sociomedical or psychological research, or any other type of social research. The absence of such rules evidently facilitated Alexander Graham Bell's original genealogical research on deafness, based on records available from the 1790 census onwards (Bruce, 1975).

What is new then is not the occurrence of privacy concerns in social research, but rather their incidence and character. Social scientists,

including those who have been traditionally uninterested in field research, have become more involved in identifying social problems and testing possible solutions through field studies. This increase in the policy relevance of research generates conflict with some policy makers simply because a new standard--higher-quality empirical data--is being offered as a substitute for a more traditional emphasis on anecdote and expert opinion. The increased contact between social scientists and individuals who are unfamiliar with their methods, objectives, and standards is almost certainly a cause of increased discord, including argument about privacy. Finally, the larger research efforts typically involve a variety of interest groups and commentators. The interaction of research sponsors, auditors, journalists, and groups of research participants with opposing views on the value and implications of the research complicates matters. In this setting, privacy arguments may distract attention from far more important issues; they may be entirely specious simply because reporting is inaccurate; or they may be legitimate but intractable because the standards of interest groups differ remarkably.

Corruption of the Principle

It does not take much imagination to expect that, at times, a confidentiality principle will be used honorifically. In the best of these instances, the appeal to principle is pious but irrelevant--that is, there is no real threat to individual privacy or to confidentiality of records. At worst, the appeal is corruptive, dedicated not to preserving individual privacy but to assuring secrecy that runs counter to the public interest.

In either case, social research and especially the evaluation of social reforms are likely to be impeded. Lobenthal (1974), for example, reports that in designing evaluative research on correctional facilities:

> Even many [correctional] program personnel from whom we sought information rather than advice withheld their cooperation. There was, for example, a sudden solicitude about clients' rights to privacy and ostensible concern with the confidentiality of records. When an elaborate protocol was worked out to safeguard confidentiality, the data we requested were still not forthcoming. (p. 32)

Similarly, the privacy issue has been used to prevent legitimate evaluations of some drug treatment programs in Pennsylvania, where records were destroyed despite immunity of record identifiers from subpoena under the

1970 Drug Abuse Act. It has been used to prevent evaluation of manpower training programs in Pittsburth and evaluation to mental health services programs in southern California. It has been used to argue against the System Development Corporation's studies of integration programs in the New York City school system, despite the fact that children who responded to inquiries would be anonymous. These episodes do not represent the norm, of course. They do represent a persistent minority event.

Little vignettes at the national level are no less noteworthy, though the reasons for impertinent appeals to privacy differ a bit from the ones just described. For example, according to Boeckmann (1976), Senate subcommittee members used the privacy issue as a vehicle for discrediting researchers during hearings on the Negative Income Tax Experiment. She suggests that the action was part of a drive to bury the idea of a graduated income subsidy program. More generally, the privacy issue has been a convenient vehicle for assaulting national research that could threaten political interests, and for getting votes. Mueller (1976), for example, argues that former President Nixon's support of the Domestic Council on Privacy, the Privacy Act, and theories of executive privilege did what it was supposed to do--focus public attention on matters other than war. That both uses are persistent but low-frequency events is evident from similar experiences in Norway, Sweden, and Denmark, as well as in the United States (Boruch & Cecil, 1979).

The most predictable adulteration of principle occurs before each U.S. population census, when ritualistic assault competes with thoughtful criticism for public attention. To Charles W. Wilson, a former chairman of the House Subcommittee on Census and Statistics, for example, much of the controversy over the 1970 Census was deliberately fomented by colleagues interested less in privacy than in votes, and by journalists moved less by the need for balanced reporting than by the need to generate provocative stories. Further, the evidence used in attacks on the census was often misleading.

> Reference was continually made to a total of 117 [Census] questions despite the fact that this total could be obtained only by adding all the different inquiries on the forms designed for 80% of the population, those for 15%, and those for 5%. A number of the questions appeared on one form only, and the maximum number of

> questions for any individual was actually less than 90. The question on whether the bathroom was shared continued to be distorted into the much more interesting version "With whom do you share your shower?" (Eckler, 1972, p. 202)

Similarly, in House Subcommittee Hearings, "One witness who had been scheduled to appear in support of legislation, proposed by Congressman Betts to restrict the 1970 Census, admitted that he had learned from earlier witnesses that his prepared statement was incorrect" (Eckler, 1972, p. 204).

An agency's refusal to disclose data on even anonymous individuals, under false colors of privacy, is of course not a new problem, nor is it confined to the social science arena. Its origins, in the United States at least, date from the reluctance of the Massachusetts Bay Colony to disclose either statistical information on mortality rates or records on the death of identified individuals, for fear of jeopardizing their project (Cassedy, 1969). The data, if disclosed, would presumably have made the colony much less attractive a prospect for volunteer colonists and for its conscientious sponsors. A similar reluctance appears to underlie the distortion of fatality and accident rates published by commercial contractors for the Alaska pipeline (see the New York Times, 7 August 1975). Institutional self-protection of the same type has hampered the efforts of biomedical researchers to understand the causes of the Thalidomide tragedy: the pharmaceutical company has refused to disclose its data on test subjects in statistical summary form or otherwise. The idea is implicit in the refusal of the Philadelphia public school system, during 1975-76, to disclose data on minority groups to the U.S. Office of Civil Rights on the grounds of student privacy, though OCR required only statistical summary data. It is transparent in at least one court case involving a school's efforts to resist, on Privacy Act grounds, the sampling of anonymous students by researchers who were interested in the racial biases that my underlie diagnosis of maladjusted and emotionally disturbed youths [Privacy Journal, 1977; Lora v. Board of Education of City of New York (74 F.R.D. 565)].

There are, at times, good administrative and political reasons for an agency's refusal to disclose statistical records to a researcher or to permit researcher access to individuals. Though we may be unable to

subscribe to those reasons, it is not in our interest to confuse the reasons for refusing disclosure with the issue of individual privacy. It is reasonable to anticipate that controversy will be instigated for purposes other than those advertised, even if we can offer no general advice here on preventing dispute. And we can offer partial solutions to one problem.

7. Public Interests and the Quality of Evidence in Public Policy

Reasoning from information is often not easy. And if the information is of an unfamiliar sort, as statistical data are for many, the task is more difficult. Perhaps more important, the unfamiliarity makes it difficult to persuade others that the information can indeed be useful and ought to be valued at least as much as experience and anecdote.

As one might suspect, the problem is an old one. No formal history of public interest in evidence for policy purposes has been written. But it should come as no surprise that arguments about the matter are as old as recorded efforts to consolidate for public policy. Consider, for instance, John Graunt's (1662/1973) Natural and Political Observations on Bills of Mortality, a progenitor of modern tracts on policy statistics. The conclusionary chapter poses a question:

> "It may be now asked to what purpose tends all this laborious bustling and groping? To know the number of . . . people, fighting men, teeming women, what years are fruitful, what proportions neglect the Orders . . ." and so forth (p. 71).

Graunt makes no bones in his first response:

> "To this I might answer in general by saying that those who cannot apprehend the reason of these enquiries are unfit to trouble themselves to ask them" (pp. 71-72).

His second reason places him among many contemporary statisticians--

> ". . . it is much pleasure in deducing so many abstract and unexpected inferences."

And his third is more politic--

> ". . . the foundation of this honest and harmless policy is to understand the land and the hands of the territory to be governed according to all their intrinsic and accidental differences. . .by the knowledge whereof trade and governmnet may be made more certain and regular. . .so as trade might not be hoped for where it is impossible. . .(all) necessary to good, certain, and easy government. . ." (p. 73-74).

Graunt's later remarks make it clear that he thinks it is in government's interest to pay attention to statistics. Bu he is not at all convinced that there's any reason for disclosing the data to the general public (see the section on Access).

Despite such early efforts, the history of "evaluation," as a formal and sustained interest of government, is embarrassingly brief. The problems engendered by collecting high-quality evidence of course are not. This has some implications for the accuracy of our views of contemporary progress and products of evaluation.

Progress is Slow

In the United States, exploiting high-quality information about social problems has been of episodic, rather than sustained, national interest, and progress is more typically sluggish than not. For instance, it was not until the 19th century that the country systematically confronted major flaws in the decennial census, longer to rectify them, despite the periodic recognition of problems in Europe, and elsewhere as early as the 14th century. Naturally, rectification was stimulated by crisis. In the 1840 census, black residents of entire towns were enumerated as insane by interviewers with more political zeal than integrity (Regan, 1973). The remedial action, appointment of census directors and regular staff partly on the basis of merit rather than on politics alone, helped. But another 80 years passed before the Census Bureau initiated a program of routine side studies on the quality of census data.

Similarly, there were some interesting efforts by statisticians to assay the effect of law or other social intervention on statistics after the Civil War. Calkins (1890-91), for instance, published a careful article assaying the effect of England's first major public health act on mortality, managing to detect and correct computational errors in the process. Indeed, he copies earlier work by Farr and does a little cost benefit analysis of the law estimating the value of human life at $770.36 per head (1890 U.S. dollars of course). Yet the practice of evaluation of any sort much less cost/benefit evaluation does not appear to have been a routine requirement of legislation for another 70 years. The earliest randomized field tests of medical regimens were undertaken in the early 1930's. But well-designed randomized field

tests did not become common until the 1960's, and a fair number of poorly designed evaluations continue to be carried out (Cochran, 1976). The proportion of such trials reported in medical journals has increased from 9% in 1963 to 46% in 1978; of the remainder, most appear to use no controls at all (Chalmers & Schroeder, 1979).

The execution of randomized field tests of nonmedical regimens has not been especially routine either. Interest in experimental tests of social services programs which appeared in the 1930's (notably in hygiene) failed to continue, though it was rekindled in the 1970's (Riecken & Boruch, 1978). Judging from Braham (1979), efforts to mount scientific tests of weather modification methods can be traced to 1946, despite a long history of rituals designed to produce rain. He suggests further that the first fifteen years of such tests did not lead to much useful information about seeding but did yield development research tools useful in the tests. The same development and lack of clear results characterized tests of educational programs during the 1960's and 1970's as well.

This inconsistency is not peculiar to medicine or the social sciences. The history of technology, for example, suggests that 25 bridges collapsed each year following the Civil War, but high-quality tests and adherence to structural standards did not become routine for another 60 years. The current renewed interest, among engineers if not the public, in bridge failures during the middle 1970's suggests that attention to quality control 30 years ago was rather too modest. The extent of interest in quality of evidence in any human enterprise, and especially in social program evaluation, is recognized only occasionally. This often engenders naive opinion about the process, and that, in turn, has some implications for government posture.

Pockets of Interest

The occasional intensive efforts of citizen's groups to collect reliable data bearing on social problems is traceable at least to the early 1800's. Lecuyer and Oberschall (1968) for instance, attribute the appearance of local statistical societies in England during the 1830's to the general interest in social reform. The societies apparently organized private applied social research of a quantitative sort to investigate health, working conditions of

the poor, and education. Interviewers were hired and sent door to door. Similar groups appeared in post-revolution France and in Germany during the middle 1800's. Lecuyer-Oberschall identify Paris's efforts to abate prostitution as an illustration of an early municipal evaluation. For the United States, Kaestle and Visnovski (1980) suggest that it is "no accident that the appearance of the first systematic school of statistics coincides with the educational reforms of the late 1830's and 1840's. The data were a crucial tool for reformers in their public relations efforts" (p. 10). The spirit of the enterprise reappeared in England in 1937 with the mass observation movement, propelled by the belief that anyone can make systematic inquiries about social phenomena (Barnes, 1979, p. 51-52).

A similar spirit is reflected in contemporary private surveys of voluntary organizations such as the League of Women Voters, Common Cause, and the like. The more technical varieties include the Stanford Workshops on Political and Social Issues (SWOPSI) in the physical and engineering sciences, the various Committees of the Assembly of Behavioral Sciences of the National Academy of Sciences, and others.

Ways of Knowing and Inept Evaluation Design

At least part of the variability of general interest in evidence is traceable to an embarrassment of riches. There are lots of ways of knowing, of apprehending information, and lots more ways of reasoning from the information.

Conflict between one stereotypical way of knowing and contemporary scientific method is exemplified by the battery additive case in 1951-54. In that instance, a chemical manufacturer claimed that one of his products increased life of storage batteries significantly, despite the National Bureau of Standards' tests on related compounds and the negative tests results. NBS was eventually asked to test the product, providing evidence of charges of fraud brought against the manufacturer by two government agencies. The NBS eventually reported that the additive had no detectable effects despite claims by the manufacturer, testimonials from trucking companies, and other manufacturers. The ensuing battle pitted small business against regulatory power of government, and more important here, evidence of a less formal sort against evidence obtained in more systematic fashion.

The seriousness of the debate was reflected partly by the Secretary of Commerce's asking for the resignation of NBS's director. The tone of at least one side of the argument is reflected in the Secretary's characterizing himself as "a practical man" rather than a man of evidence. Practicality was also espoused by the Chairman of the Senate Committee on Small Business who questioned the director: "The simple truth of the question is that if a good hardfisted businessman has used the product in a fleet of motors. . .and places orders month after month, what is the matter wiht him? Or otherwise, what is the matter with the Bureau of Standards' Test?" (p. 159). The director's understated response emphasized the controlled conditions required for scientific inference and cited Sinclair Lewis's Arrowsmith to illustrate. That appears to have been necessary but not sufficient to the eventual NBS victory.

The more dramatic examples of inept evaluation design have occurred in medicine, where medical or surgical remedies, adopted on the basis of very weak evidence, have been found to be of no use at best and to be damaging to the patient at worst. Case studies are not too difficult to find.

For instance, the so-called frozen stomach approach to surgical treatment of duodenal ulcers, for example, was used by a variety of physicians who imitated the technique of an expert surgeon. Later well-designed experimental tests showed prognoses were good simply because the surgeon who invented the technique was good at surgery and not because his innovation was effective. It provided no benefit over conventional surgery (Ruffen et al., 1969).

Prior to 1970, anticoagulant drug treatment of stroke victims had received considerable endorsement by physicians who relied solely on personal, observational data for their opinions. Subsequent randomized experimental tests showed not only that a class of such drugs had no detectable positive effects but that they could be damaging to the patients' health (see Hill et al., 1960, and other examples described in Rutstein, 1969).

There are localized examples too of course. Consider, for instance, a recent Science (Vol. 107, 1980, p. 161) article on the use of snake venom. A Florida physician claimed 20% cure rates of patients treated for multiple

sclerosis with venom. The physician received enough publicity to force the FDA to give it attention. The Food and Drug Administration sponsored a workshop to determine if the evidence justified the design and execution of controlled clinical trials. The main conclusion seems to be that the evidence is weak, and moreover that multiple sclerosis (one of the diseases for which cures were claimed), "follows such an erratic path that it's impossible to attribute improvements to any therapy without double blind studies." The evidence for MS people was not sufficient to override tests of other options.

There have been some recent efforts to characterize this problem statistically. One such approach has been to illustrate a declaration that a program or regimen is successful depends on quality of the design. Consider, for instance, Gordon and Morse's (1975) review of published evaluations of social programs. Their appraisal suggests that the probability of an evaluator winding up with a declaration that the program was a "success" based on a poor design is twice the probability based on good designs. Chalmer's (1972) analysis of a small sample of medical investigations on estrogen therapy of prostate carcinoma suggests that enthusiastic support of the therapy was almost guaranteed when the experiment was poorly designed. Improvement takes time. And there has been an improvement at least in the sense that better designs are being used more frequently. For instance, Chalmer and Schroeder's (1979) estimates of the proportion of experiments reported in the New England Journal of Medicine suggest that there has been a five fold increase in the number of studies employing randomized controls over a 25 year period to 1978. A similar analysis of studies appearing in Gastroenterology suggests that the fraction of excellent ones has increased from 5% to about 30% during 1953-1967. Similar problems are alleged to have affected the food industry. According to Samuel Epstein (National Academy of Sciences, 1974, p. 221), in 1967, 50% of all petitions submitted to the FDA in support of food additives were rejected. . .because of incomplete, inadequate, or non-specific data" (p. 221). (I have not been able to locate more recent estimates.)

Choice, Approximation, and Compromise

The need to choose between acquiring statistical information whose character is well understood and obtaining information of a less formal sort occurs often. In evaluative research, for instance, it emerges in debates about whether to invest in randomized field tests rather than in less expensive designs that yield more ambiguous information. It appears in debates about whether to mount designed surveys or to settle for a New Yorker essay based on a quick site visit. It is implicit in contemporary arguments over the proper balance of Service Delivery Assessments (fast turnaround studies) and more elaborate research. The arguments often pit manager against technologist, substantive expert against statistician, approximators against purists.

The problem is an ancient one, judging from Rabinovitch's (1973) little monograph on statistical inference in medieval Jewish literature. In discussing the idea of variability and sampling in the talmud and mishnah, he describes a second century rabbinical argument over the appropriateness of taking a systematic sample, of olives say in the interest of judging worth of crop for tithing, rather than an informal one--grabbing a convenient handful and making a declaration about worth. One result of debate appears to have been that "only in matters of lighter consequences, for example, prohibitions that are of rabbinic but not biblical origin, may one assume that perfunctory mixing gives an accurate sample." (p. 83). Roughly speaking, the rabbis' judgment was that approximation is then permissible for management purposes. It is not seemly if the demand comes from a durable and important source, such as God.

In engineering, similar tension is reflected in other ways. Borg (1962) for instance suggests that structural engineering evolved into two camps with less than cordial relations: engineering elasticity and strength of materials. The first counted the theoreticians and mathematicians among its members. The second comprised builders, crushers, and benders, engineers with a taste for the concrete so to speak. Something of the same spirit characterizes the split between experimental physics and its theoretical sister, fluid dynamics, and other fields. The gap in statistics is wide enough to concern the professional community, judging from the presidential address at the International Statistical Institute's Warsaw meetings. And it

characterizes at least a few bitter struggles in economics during the 1930's (Strotz, 1978). Settling on the appropriate level of precision or at least developing a rationale for it has been no easier for the historical demographer, judging from Hollingsworth (1968). The fact that the decision must rest on still earlier ones which might not be explicit makes matters more difficult. Early Chinese censuses were confined largely to cultivators, able bodied men, and taxpayers. "About the total population, the soverign did not wish to know" (Jaffe, 1947, p. 309).

The result in engineering at least sometimes takes a form similar to one taken in the social sector, though it is more formal than the latter. So, for example, standards for the classification of geodetic control have been developed and pinned to functional uses of the information. Local geodetic surveys are subject to less rigorous standards of precision than are scientific studies and metropolitan area surveys. The idea of tolerance bands of this sort characterize most engineering disciplines of course and the product depends on the use to which the bearing, strut, detonation timer, and so on is put. The depth to which the idea has penetrated in the social sector is not great. It is present in any formal statistical design or statistical power analysis. It is not evident in regulations that require uniform evaluation methods at local and state level, though, and there appear to have been no systematic treatments of the usefulness of broadening tolerance limits, numberical or otherwise, in dealing with local enterprise.

Language

With customary style, John Kenneth Galbraith announced that "a certain glib mastery (of the language of economics) is easy for the unlearned and may even be aided by a mildly enfeebled intellect." The language, like the vernacular of other social sciences may invite seduction because it deals with human affairs. But other aspects of scientific vernacular are interesting, and the physical sciences are not entirely immune to the problems it engenders. These features include the creation of new, official meanings for existing words, causing confusion among the profane. If the scientist, especially the social scientist, seeks to avoid the problem by inventing new words, then lexicographic assault may follow. The confusion mounts when the new words are popularized mistakenly in the press or by public representatives.

To be sure, emerging areas of inquiry such as evaluation are usually characterized by a good deal of lexical ambiguity. Glass and Elliott (1980) rummage through contemporary papers to find evaluation defined as applied sciences, systems management, decision theory, assessment of progress toward goals, description or portrayal, and rational empiricism. In our own investigations (Boruch & Cordray, 1980), we have interviewed a director of research who announced his office did no evaluation, and his boss, at the deputy secretary level, who announced that everything they are responsible for is evaluation. We encountered Congressional staffers who, in criticizing research or evaluation, fail to distinguish among evaluation, research, development, and monitoring. We also talked to support agency staff members who eschew the word evaluation entirely, preferring instead simply to specify what question is answered by the process: Who is served? How well are they served? How much does it cost? And what are the effects of service? The phrases invented by academicians to clarify are sometimes remarkably effective in consolidating a variety of related themes under a single rubric. The less durable ones confuse and it is difficult "to praise famous coiners of new words and the happy nomenclators that begat them" (Howard, 1979, p. 153) if the new ones are no better than the old. The student is offered "formative" evaluation instead of trouble-shooting or development, "summative" evaluation instead of estimating program effects, and "meta-analysis" instead of synthesis or combining estimates of effect. These and other new phrases have become potois in much less than ten years. There are still many evaluators who try to speak English, however.

The adoption of some of these words by politicians and journalists has its parallel in the adoption from other disciplines of phrases that are aurally attractive and equally vague. The phrase "representative sampling" for instance, has no formal definition in mathematical or applied statistics. In the nonscientific literature, for example, it is used to imply that given sample has a sort of seal of approval or a vaguely scientific dignity. It is used to indicate a miniature of a population, to suggest that extreme or varied units have been examined (Kruskal & Mosteller, 1979a). In the nonstatistical scientific literature, it is used

as a less forbidding synonym for random sampling (Kruskal & Mosteller, 1979b). I expect that the word experiment" is used in at least as many ways. The word was appropriated from simple language by statisticians. Unless told otherwise, the latter would expect the thing to be randomized, and it is now used to lend an aura of scientific legitimacy to the process of merely trying things out in laboratory and field settings.

There are also words that have become popular by mistake. Their misuse is more pleasing than proper use, and in any event, it is hard for the non-scientist to understand the correct definition. Howard's (1979) catalog of words of this ilk is fascinating: "Quantum jumps" are not very big ones as its users usually imply; rather, they are exceeding small transitions from one energy state to another. He suggests, incidentally, that the term's abusers be made to walk the Planck, because they have got hold of the wrong end of the quark. Feedback too has gotten appropriated inappropriately. Geriatric implies health for the aged; its reference to the long of tooth is incomplete.

The problem is not a new one of course. In social statistics, at least, it's recorded history dates at least to C. F. Pidgin's (1890-91) efforts to popularize statistics in a seemly fashion. He objected vigorously to the gobbledygook invented to praise, to obscure, and especially attack: "Now we have statistical roorbacks (supplementing the literal variety) and neither the politicians nor the people understand them (p. 109)." He also made a plea for simple summary and homely comparisons, echoed 85 years later by the New York Times, e.g., "How does the risk the FDA has moved against compare with the risk of breathing normal polluted air in Manhattan?" (National Academy of Sciences, 1974, p. 27).

The lexical difficulties are tedious, frustrating, and unnecessary at times. In short, they are normal. The puzzling part is why we do not expect them and have no better ways to deal with them.

8. Use of Randomized Experiments in the Physical Sciences

During a recent oral examination, a social science student suggested that because randomized tests are not often used in the physical sciences and engineering, one ought to be suspicious about their utility. The premise, rarity of experiments in the area, is not unreasonable when one considers that few undergraduate science courses stress the topic. But it does not reflect reality well.

Consider, for example, recent research on weather control. Over the past 25 years, both randomized experimental tests, as well as quasi-experiments, have been run to determine whether the introduction of silver iodide crystals into the air will under certain conditions increase the probability of precipitation. As in the social sciences, the incidence of non-randomized experiments exceeds that of randomized trials. The work does involve the physical sciences since weather dynamics, chemistry, as well as some knowledge of the natural sciences such as atmospherics. This is also a nice illustration of a research endeavor in which the distinction between physical sciences and natural sciences ceases to be meaningful. So-called Grossversuch 3, which lasted seven years and ended in 1963 is among the largest of weather experiments. The unit of randomization in the experiments was a 24 hour period. Each period was randomly assigned to an activity of seeding clouds or to an activity of not seeding clouds, conditional on prior predictions about whether thunder storms with hail were to be expected on the day in question. The experiments were conducted on the southern slopes of the Alps in Switzerland and Italy (see Neyman, 1977, for example).

Related efforts include the National Hail Research Experiment, undertaken in Colorado and Nebraska in 1972-1974 hail seasons (see Crow et al., 1977). The experimental unit was the declared hail day, determined on the basis of radar reflectivity data. "A random 50/50 choice to seed or not to seed was applied only to the first day of a sequence of one or more hail days; subsequent days in a sequence were given alternating treatments." Treatment consisted of seeding clouds with silver iodide crystals using rockets, pyrotechnic flares, etc. Response variables were hail size (smaller circumference than control days), rain mass, and others. Apparently seeding had no effect at the 10% level of significance.

The randomized trials describe by Braham (1979) have been subjected to intensive independent scrutiny and secondary analysis. The material given earlier in Section 2 suggests that there are parallels between the operational problems in meteorological experiments and experimentsl tests of social programs. The Colorado trials described by Elliott et al. (1978) exhibit similarities as well.

A second broad category of randomized experiments in the physical sciences involves tests of material strength of materials as a function

the shape of the material, its composition, and other factors. Here again, the use of randomized experiments is less frequent than the use of non-experiments. However, one can find controlled studies of, for example, the effect of fibre diameter on the fatigue strength of metal composites making up the fibres. In civil engineering research, it is not difficult to find randomized experiments in which the hardening time and characteristics of the thickening process of cement are examined as a function of temperature, pressure, and other physical properties of the cement. The unit of randomization is a sample from a batch, several units being extracted from each batch in order to make up replications. In chemistry, the light fastness of dye bases has been explored using randomized experiments with chemical composition of the dye base as a treatment variable, and a stable element of the dye base as a blocking variable. The problems of designing efficient experiments in chemical processing have been sufficient to produce a substantial body of literature on optimization, by V. V. Fedorov at the University of Moscow among others.

More generally, it's not difficult to find major technical reports on randomized experimental designs in the physical sciences and engineering, issued by, among others, the National Bureau of Standards and the General Electric Corporation. Each has research laboratories which have produced reports on fractional factorial designs. The Office of Aerospace Research at Wright Patterson Air Force Base in Ohio has issued important reports on complex fractional factorial designs.

Finally, there are at least a half dozen textbooks available on experimental design in the engineering sciences, especially industrial engineering and related areas. Books by Brownlee, Daniel (1976), Davies (1971), and Chew (1958) are notable. Still more generally, a sizable number of individuals who've made distinctive contributions to applied statistics over the past 20 years have done so through their involvement with applied research in the physical sciences. This includes, for example, G. E. P. Box (Hunter & Box, 1965) and H. Scheffe (1958), as well as individuals who are better known for their work in agricultural research, such as Youden and Kempthorne. In fact, some major areas of experimental design have grown primarily out of work in the industrial and engineering sector: fractional factorials, weighing designs and specialized designs for understanding chemical mixtures

and the influence which externally manipulated factors around them. The work is reported regularly in journals such as Technometrics (e.g., Webb, 1973), and Biometrics (e.g., Davies & Hay, 1950).

Footnotes

1. I am grateful to the support of the National Science Foundation (DAR 7820374) for support of work on evaluative methods in the social sciences, and to the National Institute of Education (NIE-G-79-0128) for support of work on evaluation in education. Portions of this paper have been presented at the University of Jerusalem in June 1980 and at the U.S. General Accounting Office. William H. Kruskal kindly provided suggestions on an earlier draft.

2. The failure of a suspension bridge over the Main River (France) in 1850 also provides a nice illustration of malformed federal regulations. The amplified vibration was caused by the Eleventh Light Infantry's cadence march. All perished in the ensuing bridge collapse. Daumas and Gille (1968/1979) report that general orders were issued that infantry break step on bridges as a consequence. Because bridge type was not specified, infantry broke step on all bridges.

3. Samuel Johnson did not pussyfoot around palmistry either. His dictionary defined it as "the cheat of foretelling the future by the lines of the palm" (Johnson, 1955/1979).

4. The clever exploitation of casualty statistics is occasionally matched by clear contemporary reporters. For instance, Henry Kissinger's autobiography claims that his decisions about invading Cambodia were justified partly on account of the (anticipated) continuous drop in American casualty rates following invasion. William Shawcross neatly assaults K. in a footnote, citing the cyclical character of casualties in Viet Nam and the declining secular trend underlying the cycle as a competing, more plausible explanation for the decline, a trend attributable mainly to the gradual withdrawal of troops from combat areas (Harper's, November 1980, pp. 35-44, 89-97).

5. The effort to secure the original specimens appears to have been vigorous. The denouement seems not to have impaired much the installation of Lysenko's beliefs about inheritance of acquired characteristics in Russia during the 1950's (Zirkle, 1954).

6. Excerpted from Boruch and Cecil (1979).

REFERENCES

Accum, F. A treatise on adulterations of food, and culinary poisons, exhibiting the fradulent sophistications of bread, beer, wine, spiritous liquors, tea, coffee, cream, confectionery, vinegar, mustard, pepper, cheese, olive oil, pickles, and other articles employed in domestic economy, and methods of detecting them. Philadelphia: Abr'm Small (Mallinkrodt collection), 1820.

Altman, L. K. 2 TB vaccines found ineffective. New York Times, January 20, 1980, pp. 1, 44, 45.

Amberson, J. B., McMahan, B. T., & Pinner, M. A clinical trial of soncrysin in pulmonary tuberculosis. American Review of Tuberculosis, 1931, 24, 401.

Angle, P. M. The minor collection: A criticism. Atlantic Monthly, 1929. Reprinted in R.W. Winks (Ed.), The historian as detective: Essays on evidence. New York: Harper & Row, 1968, pp. 127-141.

Barnes, J. A. Who should know what? Social science, privacy, and ethics. New York: Cambridge University Press, 1979.

Barry, B. A. Errors in practical measurement in science, engineering, and technology. New York: Wiley, 1978.

Benjamin, B. Quality of response in census taking. Population Studies, 1966, 8(3), 288-293.

Berger, R. Letter to the editor. Earthquake resistant buildings. Science, 1980, 207, 478.

Bernard, J. S., & Bernard, L. L. Origins of American Sociology. New York: Thomas Y. Crowell, 1943.

Boeckmann, M. E. Policy impacts of the New Jersey Income Maintenance Experiment. Policy Sciences, 1975, September.

Borg, S. Fundamentals of engineering elasticity. Princeton, J.J.: Van Nostrand, 1962.

Boring, E. G. The nature and history of experimental control. American Journal of Psychology, 1964, 67, 573-589.

Boruch, R. F., & Cecil, J. S. Assuring confidentiality of social research data. Philadelphia: University of Pennsylvania, 1979.

Boruch, R. F., & Cordray, D. S. (Eds.). An appraisal of educational program evaluations: Federal, state, and local agencies. Evanston, Ill.: Northwestern University, 1980.

Boruch, R. F., Wortman, P. M., & Cordray, D. S. (Eds.) Secondary analysis of social program evaluations. San Francisco: Jossey-Bass, 1980.

Boruch, R. F., & Gomez, H. Power theory in social program evaluation. In L. Datta & R. Perloff (Eds.), Improving evaluations. Beverly Hills, Ca.: Sage, 1979.

Borus, M. E. Measuring the impact of employment related social programs. Kalamazoo, Mich.: W. E. Upjohn Institute for Employment Research, 1979.

Bowers, W., & Pierce, G. The illustration of deterrence in Isaac Ehrlich's research on capital punishment. Yale Law Journal, 1975, 85, 187-208.

Box, G. E. P. Use and abuse of regression. Technometrics, 1966, 8, 625-629.

Braham, R. E. Field experimentation in weather modification. Journal of the American Statistical Association, 1979, 74(365), 57-104.

Bunker, J. P., Barnes, B. A., & Mosteller, F. (Eds.). Costs, risks, and benefits of surgery. New York: Oxford University Press, 1977.

Burrows, M. The dead sea scrolls. New York: Viking, 1955.

Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H., & Ware, J. H. Randomized clinical trials: Perspectiveness on some recent ideas. New England Journal of Medicine, 1976, 295, 74-80.

Calkins, G. N. Some results of sanitary legislation in England since 1875. Journal of the American Statistical Association, 1890-91, 2, 297-303.

Callahan, R. E. Education and the cult of efficiency. Chicago: University of Chicago Press, 1962.

Campbell, D. T., & Boruch, R. R. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C.A. Bennett and A.A. Lumsdaine, Evaluation and experiment. New York: Academic Press, 1975, pp. 195-297.

Caplan, R., & Caplan, G. Psychiatry and the community in nineteenth century America. New York: Basic, 1969.

Carter, L. J. Job protection for "whistle blowers" being tested. Science, 1980, 207, 1057.

Cassedy, J. H. Demography in early America: Beginnings of the statistical mind, 1600-1800. Cambridge: Harvard University Press, 1969.

Chalmers, T. C. A challenge to clinical investigators. Gastroenterology, 1969, 57(6), 631-635.

Chalmers, T. C., & Schroeder, B. Controls in Journal articles. New England Journal of Medicine, 1979, 301, 1293.

Chalmers, T. C., Block, J. B., & Lee, S. Controlled studies in clinical cancer research. New England Journal of Medicine, 1972, 287, 75-78.

Chambers, L. W., West, A., Ho, C., Hunt, G., Lawlor, B., & Power, R. Health outcomes of patients in the St. John's randomized trial of the family practice nurse. Paper presented at the meeting of the Society for Epidemiologic Research, Seattle, Washington, June 15, 1977. (Available from the authors, Division of Community Medicine, Memorial University of Newfoundland.)

Chambers, S. P. Statistics and intellectual integrity. Journal of the Royal Statistical Society, Series A, 1965, 128, 1-15.

Chapin, F. S. The problem of controls in experimental sociology. Journal of Educational Sociology, 1931, 4(9), 541-551.

Chapin, F. S. Design for social experiments. American Sociological Review 938, 3(6), 786-800.

Chapin, F. S. Experimental designs in sociological research. New York: Harper, 1947.

Chew, V. (Ed.). Experimental designs in industry. New York: Wiley, 1958.

Clemens, S. L. The autobiography of Mark Twain. New York: Harper (originally published 1917), 1959.

Cochran, W. B. Early developments in comparative experimentation. In D. B. Owen (Ed.), On the history of probability and statistics. Basel: Marcel Dekker, 1976, pp. 1-27.

Cochran, W. G., Diaconis, P., Donner, A. P., Hoaglin, D. C., O'Conner, N. E., Peterson, O. L., & Rosenoer, V. M. Experiments in surgical treatment of duodenal ulcer. In J. P. Bunker, B. A. Barnes, & F. Mosteller (Eds.), Costs, risks, and benefits of surgery. New York: Oxford University Press, 1977, pp. 176-197.

Cook, R. D., & Holschuh, N. Comment. Journal of the American Statistical Association, 1979, 74(365), 68-70.

Cremin, L. A. The transformation of the school: Progressivism in American education, 1876-1957. New York: Vintage, 1964.

Crow, E. L., Long, A. G., Dye, J. E., Mielke, P. W., & Ulrich, C. W. Primary statistical evaluation of the National Hail Experiment: Randomized seeding experiment 1972-74. Proceedings of the Sixth Conference on Inadvertant and Planned Weather Modification. Boston: American Meterological Society, 1977, pp. 150-153.

Cunliffe, S. V. Address of the president. Journal of the Royal Statistical Society, 1976, 139 (Part 1), p. 1-16.

Cutler, J. L., Ramcharan, S., Feldman, R., Sieglaub, A. B., Cambell, B., Friedman, G. D., Dales, L. G., & Collen, M. F. Multiphasic checkup evaluation study: I. methods and populations. Preventive Medicine, 1973, 2, 197-206.

Dain, N. Concepts of insanity in the United States: 1789-1865. New Brunswick, N.J.: Rutgers University Press, 1964.

Daniel, C. Applications of statistics to industrial experimentation. New York: Wiley, 1976.

Daumas, M., & Gille, P. Roads, bridges and transports. In M. Daumas (Ed., translated by E. Hennessy), A history of technology and progress. Volume III, 1725-1860. New York: Crown, 1968/1979, pp. 235-257.

Davies, O. L. (Ed.). The design and analysis of industrial experiments. San Francisco: Hafner, 1971.

Davies, O. L., & Hay, W. A. Construction and uses of fractional factorial designs in induatrial research. Biometrics, 1950, 6, 233-249.

Davis, R. M. Social research in America before the Civil War. Journal of the History of the Behavioral Sciences, 1972, 8, 69-85.

Dawkins, S. M., & Scott, E. L. Comment. Journal of the American Statistical Association, 1979, 74(365), 70-71.

Dearnaley, E. J., & Warr, P. B. (Eds.). Aircrew stress in wartime operations. London: Academic Press, 1978.

de Santillana, G. The crime of Galileo. Chicago: University of Chicago Press, 1955.

Diet-Heart Feasibility Study Group. The Diet-Heart Feasibility Study, Circulation (Supplement 1), 1968, 1-428.

Dodd, S. C. A controlled experiment on rural hygiene in Syria (Soc. Sci. Ser. No. 7). Beirut: 1934, 128.

Duffy, J. A history of public health in New York City: 1866-1966. New York: Russell Sage, 1974.

Don & Bradstreet. The business failure record. New York: Dun & Bradstreet, Inc., 1979.

Dyck, F. J., et al. Effect of surveillance on the number of hysterectomies in the Province of Saskatchewan. New England Journal of Medicine, 1977, 296(23), 1326-1328.

Eckler, A. R. The Bureau of the Census. New York: Praeger, 1972.

Eisenhart, C. Realistic evaluation of the precision and accuracy of instrumental calibration systems. Journal of Research of the National Bureau of Standards, 1968, 67C, 21-47.

Elliott, R. D. Comment. Journal of the American Statistical Association, 1979, 74(365), 77.

Elliott, R. D., Shaffer, R. W., Court, A., & Hannaford, J. F. Randomized cloud seeding in the San Juan Mountains, Colorado. Journal of Applied Meteorology, 1978, 17(9), 1298-1318.

Fairweather, G., & Tornatsky, L. Experimental methods for social policy research. New York: Pergamon, 1977.

Fedorov, V. V. Regression problems with controllable variables subject to error. Biometrika, 1974, 61, 49-56.

Flaherty, D. H. Privacy in colonial New England. Charlottesville: University Press of Virginia, 1972.

Flaherty, D. H. Final report of the Bellagio Conference on privacy, confidentiality, and use of government microdata for research and statistical purposes. Statistical Reporter. May, 1978. No. 78-8, pp. 274-279.

Fleck, L. Genesis and development of a scientific fact. (Translated by F. Bradley and Thaddeus Trenn, Edited by T. Trenn and R. K. Merton.) Chicago: University of Chicago Press, 1979.

Flexner, A., & Bachman, F. P. The Gary schools: A general account. New York: General Education Board, 1918.

Florman, S. C. The existential pleasures of engineering. New York: St. Martin's Press, 1976.

Flueck, J. A. Comment. Journal of the American Statistical Association, 1979, 74(365), 77-80.

Freiman, J. A., Chalmers, T. C., Smith, H., & Kubler, R. R. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 negative trials. New England Journal of Medicine, 1978, 299, 690-694.

Friedman, L. M. The legal system: A social science perspective. New York: Sage, 1975.

Gabriel, K. R. Comment. Journal of the American Statistical Association, 1979, 74(365), 81-86.

Galbraith, J. K. Economics, peace, and laughter. New York: Houghton Mifflin, 1971.

Gilbert, J. P., McPeek, B., & Mosteller, F. Progress in surgery and anasthesia: Benefits and risks of innovative therapy. In J. P. Bunker, B. A. Barnes, & F. Mosteller (Eds.), Costs, risks, and benefits of surgery. New York: Oxford University Press, 1977.

Good, I. J. A classification of fallacious arguments and interpretations. Technometrics, 1962, 4, 125-132.

Goran, M. Fact, fraud, and fantasy. London: Yoseloff, 1979.

Graunt, J. Natural and political observations made upon the bills of mortality (1662). In P. Laslett (Compiler), The earliest classics. Germany: Gregg International, 1973.

Greenwood, E. Experimental sociology. New York: Kings Crown, 1945.

Grob, G. N. Mental institutions in America: Social policy to 1875. New York: Free Press, 1973.

Grob, G. N. The state and the mentally ill: A history of the Worcester State Hospital in Massachusetts, 1830-1920. Chapel Hill, N.C.: University of North Carolina, 1966.

Hill, A. B., Marshall, J., & Shaw, D. A. A controlled clinical trial of long term anticoagulant therapy in cerebrovascular disease. Quarterly Journal of Medicine, 1960, 29, 597-609.

Hollingsworth, T. H. The importance of the quality of the data in historical demography. Daedalus, 1968, 97(2), 415-432.

Howard, P. Weasel words. New York: Oxford University Press, 1979.

Hunter, W. G., & Box, G. E. P. The experimental study of physical mechanisms. Technometrics, 1965, 7(1), 23-42.

Ingle, D. J. Fallacies and errors in the wonderlands of biology, medicine, and Lewis Carroll. Perspectives in Biology and Medicine, 1972, 15, 254-281.

Jacobs, J. L. The GAO and the consumer: The need for better food labeling. In E. H. Kloman (Ed.), Cases in accountability: The work of the GAO. Boulder, Colorado: Westview, 1979, pp. 35-42.

Jaffe, A. J. A review of the censuses and demographic statistics of China. Population Studies, 1947, 1, 308-337.

Jerome, T. S. The case of the eyewitness: A lie is a lie, even in Latin. Chapter 9 of R.W. Winks (Ed.), The historian as detective: Essays on evidence. New York: Harper & Row, 1968, pp. 181-191.

Johnson, N. L., & Leone, F. C. Statistics and experimental design. New York: Wiley, 1964.

Johnson, S. A dictionary of the English language. New York: Arno, 1755/1979.

Kaestle, C. F., & Vinovskis, M. A. Education and social change in nineteenth-century Massachusetts. New York: Cambridge University Press, 1980.

Katz, M. B. The irony of early school reform: Educational innovation in mid-nineteenth century Massachusetts. Cambridge, Mass.: Harvard University Press, 1968.

Kelling, G. L. et al. The Kansas City police patrol experiment. In G. V Glass (Ed.), Evaluation Studies Review Annual, 1976, 1, 605-657.

Kilo, C., Miller, J. P., & Williamson, J. R. The Achilles heel of the University Group Diabetes Program. Journal of the American Medical Association, 1980, 243, 450-457.

King, G. Natural and political observations and conclusions upon the state and condition of England (1696). In P. Laslett (Compiler), The earliest classics. Germany: Gregg International, 1973.

Kirk, D. Demography. In W. H. Kruskal & J. M. Tanur (Eds.), International encyclopedia of statistics. Volume 1. New York: Free Press, 1978, pp. 136-144.

Kloman, E. H. (Ed.) Cases in accountability: The work of the GAO. Boulder, Colorado: Westview, 1979.

Knightly, P. The first casualty. New York: Harcourt, Brace, Jovanovich, 1975.

Koestler, A. The case of the midwife toad. New York: Random House, 1971.

Kolata, G. B. Is labile hypertension a myth? Science, 1979, 204, 489(a).

Kolata, G. B. Controversy over study of diabetes drugs continues for nearly a decade. Science, 1979, 203, 986-990(b).

Kotulak, R. Heart bypass success tied to experience. Chicago Tribune, March 1978, Sec. 1, p. 4.

Kruskal, W. H. Comment. Journal of the American Statistical Association, 1979, 74(365), 84-86.

Kruskal, W., & Mosteller, F. Representative sampling, II: Scientific literature excluding statistics. International Statistical Review, 1979, 47, 111-127(b).

Kruzas, A. T., & Sullivan, L. V. (Eds.). Encyclopedia of information systems and services (3rd ed.). Detroit: Gale Research Co., 1978.

Large, A. J. Weather tinkerers seek two things: Results and evidence. Wall Street Journal, September 14, 1979.

Lécuyer, B., & Oberschall, A. R. The early history of social research. In D. L. Sills (Ed.), International encyclopedia of the social sciences. New York: Macmillan and Free Press, 1968, pp. 1013-1031.

Leimer, D. R., & Lesnoy, S. D. Social security and private saving: A reexamination of the time series evidence using alternative social security wealth variables. Presented at the 93rd annual meeting of the American Economic Association, Denver, Colorado, September 6, 1980.

Leonard, W. H., & Lowery, L. F. Was there really an experiment? A quantitative procedure for verifying treatments in educational research. Educational Researcher, 1979, 8, 4-7.

Levitan, S. A., & Taggart, R. Do our statistics measure the real labor market hardships? Proceedings of the American Statistical Association: Social Statistics Section. Washington, D.C.: ASA, 1976.

Magidson, J. Toward a casual model approach for adjuting for preexisting differences in the nonequivalent control group situation. Evaluation Quarterly, 1977, 1(3) 399-420.

Marks, G., & Beatty, W. K. Epidemics. New York: Scribner, 1976.

Mausner, J. S., & Gezon, H. M. Report on a phantom epidemic of gonorrhea. American Journal of Epidemiology, 1967, 85, 320-331.

McDermott, W. Evaluating the physician and his technology. Daedalus, 1977, 106(1), 135-158.

Meier, P. The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters, & G. Rising (Eds.), Statistics: A guide to the unknown. San Francisco: Holden-Day, 1972.

Meier, P. Statistics and medical experimentation. Biometrics, 1975, 31, 511-529.

Miao, L. L. Gastric Freezing: An example of the evaluation of medical therapy by randomized clinical trials. In J. P. Bunker, B. A. Barnes, & F. Mosteller (Eds.), Costs, risks, and benefits of surgery. New York: Oxford University Press, 1977, pp. 176-197.

Middleton, T. H. The sociologese plague. Saturday Review, September 1980, 81.

Mielke, P. W. Comment. Journal of the American Statistical Association, 1979, 74(365), 87-88.

Moran, P. A. P. Problems and mistakes in statistical analysis. Communications in Statistics, 1973, 2, 245-257.

Moser, R. H. (Ed.) Diseases of medical progress: A study of iatrogenic disease. Springfield, Ill.: Charles C. Thomas, 1969.

Mosher, F. C. The GAO: The quest for accountability in American government. Boulder, Colorado: Westview, 1979.

Mosteller, F. Errors: Nonsampling errors. In W. H. Kruskal and J. M. Tanur (Eds.), International encyclopedia of statistics. Volume 1. New York: Free Press, 1978, pp. 208-229.

Mosteller, F., Gilbert, J. P., and McPeek, B. Reporting standards and research strategies for controlled trials. Controlled Clinical Trials, 1980, 1(1), 37-58.

Moynihan, D., & Mosteller, F. (Eds.) On equaltiy of educational opportunity. New York: Vintage, 1972.

Mueller, G. O. W. Prestige and the researcher's accountability. In P. Nejelski (Ed.), Social research in conflict with law and ethics. Cambridge, Massachusetts: Ballinger, 1976, pp. 111-122.

National Academy of Sciences. How safe is safe: The design of policy on drugs and additivies, Academy Forum Series. Washington, D.C.: NAS, 1974.

Nelkin, D. Controversy: Politics of technical decisions. Beverly Hills, CA: Sage, 1979.

Neyman, J. Experimentation with weather control and statistical problems generated by it. In P. R. Krishnaiah (Ed.), Applications of statistics. Amsterdam: North-Holland, 1977, pp. 1-23.

Neyman, J. Comment. Journal of the American Statistical Association, 1979, 74(365), 90-94.

Neyman, J. (Ed.). The heritage of Copernicus: Theories more pleasing to the mind. Cambridge, MA: MIT Press, 1974.

Penick, J. L., Pursell, C. W., Sherwood, M. B., & Swain, D. C. The politics of American science: 1939 to the present. Chicago: Rand-McNally, 1965, 148-160.

Perry, T. C. Slow progress in developing and implementing a national dam safety program. In E. H. Kloman (Ed.), Cases in accountability: The work of the GAO. Boulder, Colorado: Westview, 1979, pp. 243-246.

Peterson, J. C., & Markle, G. E. The Laetrile controversy. In D. Nelkin (Ed.), Controversy: Politics of technical decisions. Beverly Hills, CA: Sage, 1979, pp. 159-179.

Pidgin, C. F. How to make statistics popular. Journal of the American Statistical Asslciation, 1980-91, 2, 107-115.

Pressman, J. L., & Wildavsky, A. B. Implementation. Berkeley: University of California, 1973.

Primack, J., & von Hipple, F. Advice and dissent: Scientists in the political arena. New York: Basic, 1974.

Proudfit, W. L. Criticisms of the VA randomized study of coronary bypass surgery. Clinical Research, 1978, 26, 236-240.

Przeworski, A., & Teune, H. The logic of comparative social inquiry. New York: Wiley, 1970.

Rabinovitch, N. L. Probability and statistical inference in ancient and medieval Jewish literature. Toronto: University of Toronto Press, 1973.

Regan, O. R. Statistical reforms accelerated by sixth census errors. Journal of the American Statistical Association, 1973, 68(343), 540-546.

Resnikoff, H. L., & Wells, R. O. Mathematics in civilization. New York: Holt, Rinehart, Winston, 1973.

Rossi, P. H. The challenge and opportunities of applied social research. Presidential Address, American Sociological Association, New York City, August 18, 1980.

Roueché, B. The medical detectives. New York: Quadrangee, 1980.

Ruffin, J. N., Grizzle, J. E., Hightower, N. C., McHardy, G., Shull, I. I., & Krisher, J. B. A cooperative double-blind evaluation of gastric 'freezing' in the treatment of duodenal ulcer. New England Journal of Medicine, 1969, 281, 16-19.

Rustein, D. D. The ethical design of human experiments. Daedalus, 1969, 98(2), 523-541.

Sackett, D. I. End results analyses in a randomized trial of nurse practitioners. Research memorandum. Hamilton, Ontario. McMaster University Medical Center, Brulington Study Group, 1973.

Scheffe, H. Experiments with mixtures. Journal of the Royal Statistical Society, Series B, 1958, 20, 344-360.

Schoonover, R. M., Davis, R. S., & Bower, V. E. Mass measurement at the national bureau of standards: A revision. Science, 1980, 207, 1347-1348.

Sechrest, L., & Redner, R. Strength and integrity of treatments in evaluation studies. In How well does it work? Review of criminal justice research, 1978. Washington, D.C.: U.S. Department of Justice, Law Enforcement Assistance Administration, June 1979.

Seybert, A. Statistical annals. Philadelphia: T. Dobson & Son, 1818.

Shapiro, S., Strax, P., & Venet, L. Periodic breast cancer screening in reducing mortality from breast cancer. Journal of the American Medical Association, 1971, 215, 177-190.

Silverman, W. A. The lesson of retrolental fibroplasia. Scientific American, 1977, 236(6), 100-107.

Simpson, J. Comment. Journal of the American Statistical Association, 1979, 74(365), 95-97.

Smith, R. J. Latest saccharine tests kill FDA proposal. Science, 1980, 208, 154-156.

Sousa, G. A. The airborne warning and control system. In E. H. Kloman (Ed.), Cases in accountability: The work of the GAO. Boulder, Colorado: Westview, 1979, pp. 53-60.

Stevens, A. C. The commercial death rate. Journal of the American Statistical Association, 1890-91, 2, 186-194.

Strotz, R. Econometrics. In W. H. Kruskal & J. M. Tanur (Eds.), International encyclopedia of statistics. Volume 1. New York: Free Press, 1978, pp. 188-197.

Sullivan, J. H. Foreign affairs oversight: Role of the staff survey mission. In U. S. Senate, Subcommittee on Oversight Procedures. Committee on Government Operations. Legislative Oversight and Program Evaluation. 94th Congress, 2nd Session, Washington, D.C.: U.S. Government Printing Office, 1976, pp. 173-185.

Taeuber, C. Census. In W. H. Kruskal & J. M. Tanur (Eds.). International encyclopedia of statistics. Volume 1. New York: Free Press, 1978, pp. 41-45.

Tuberculosis Prevention Trial. Trial of BCG vaccines in southern India for tuberculosis prevention: First report. Bulletin of the World Health Organization, 1979, 57(5), 819-827.

Tyor, P. L. Segregation or surgery: The mentally retarded in America, 1850-1920. Ph.D. Dissertation, Department of History, Northwestern University, 1972.

U.S. General Accounting Office. Need to improve regulatory enforcement procedures involving pesticides (B-133192). Washington, D.C.: U.S. GAO, 1968.

Verma, R. L. Graeco-arabic medicine in medieval India - Its hakims and hospitals. Hamdard Medicus, 1977, 20, 26-40.

Watson, J. D. The double helix: A personal account of the discovery of structure of DNA. New York: New American Library, 1968.

Webb, S. Small incomplete factional designs for two and three level factors. Technometrics, 1971, 13, 243-256. (Corrigenda. Technometrics, 1973, 15, 951).

Winks, R. W. (Ed.). The historian as detective: Essays on evidence. New York: Harper & Row, 1968.

Wolins, L. Secondary analysis: Published research in the behavioral sciences. New Directions for Program Evaluation, 1978, 1(4), 45-57.

Woodward, J. A., & Goldstein, M. G. Communication deviance in the families of schizophrenics. A comment on the misuse of analysis of covariance. Science, 1977, 197, 1096-1097.

Yarmolinsky, A. How good was the answer? How good was the question? In C. Frankel (Ed.), Controversies and decisions: The social sciences and public policy. New York: Russell Sage, 1976, pp. 259-272.

Yates, F. Principles governing the amount of experimentation in developmental work. Nature, 1952, pp. 138-140.

Yerushalmy, J., Garland, L. H., Harkeness, J. T., Hinshaw, H. C., Miller, E. R., Shipman, S. J., & Zwerling, H. B. An evaluation of serial chest roentenograms in estimating the progress of disease in patients with pulmonary tuberculosis. American Review of Tuberculosis, 1961, 64, 225-248.

Zirkle, C. Citation of fraudulent data. Science, 1954, 120, 189-190.