PN-AAT-506 41577

(INTERIM REPORT)

On Impact Evaluation in Developing Countries, I.

Robert F. Boruch

PURSUANT TO: AID/OTR-147-79-106 (PPC/PDPR/SR)

PRESENTED AT: Presented in lecture form at the West African Seminar on Evaluation Techniques at Abidjan, Ivory Coast, January 7-11, 1980, and the East African Seminar on Evaluation Techniques at Nairobi, Kenya, January 14-18, 1980.

REPORT: A-92 (Second revision)

REVISED FROM: A-73

Division of Methodology and Evaluation Research
Department of Psychology
and Center for Applied Probability and Statistics
Northwestern University
Evanston, Illinois USA 60201

AID

January 10, 1980

Biographical Sketch

Professor Robert F. Boruch is Director of the Division of Methodology and Evaluation Research at Northwestern University. He holds a faculty appointment in the Departments of Psychology and Education and in the Center for Applied Probability and Statistics, the Center for Health Services Policy Research, and the Center for Urban Studies and Policy. He is a member of the Board of Directors of the Evaluation Research Society and is past president of the Council for Applied Social Research. He has served on committees of the National Academy of Sciences, including one on noise impact of Concorde landings, and one on privacy and confidentiality in social research. He is current Chairman of the Social Science Research Council's Committee on Program Evaluation.

Boruch is author of forty journal articles and editor, author, or co-author of six books. Dr. Boruch's work covers development of quantitative methods in applied research and their union with qualitative methods, development of methods of measurement, alternative approaches to assuring privacy and confidentiality of individual response in social research, research and evaluation policy.

Abstract

This report contains a summary of certain developments in evaluation, with emphasis on impact evaluations in developing countries. There are two parts to the report. Part I covers major aspects of evaluation, including types of evaluation, methods, and issues; this material was presented in lectures at the recent Agency for International Development Conferences on evaluation in the Ivory Coast (Abidjan, January 7-11, 1980), and Kenya (Nairobi, January 14-18, 1980). Part II is a compendium of references to published articles on new developments in impact evaluation.

Contents of Part I.
On Impact Evaluations in Developing Countries

1. Definition and Types of Evaluation
2. Design of Outcome Evaluations
3. Ambiguity and Bias in Estimating Program Effects
4. On Human Factors in Program Evaluation
5. Recent Developments on Uniform Structures for Reporting on Evaluations
6. Common Questions and Partial Answers

1. Definition and Types of Evaluation

The word evaluation is an indiviously simple label for a complex variety of activities and functions. USAID's Handbook, for instance, reiterates the point that home-made definitions "of evaluation" are common and varied, e.g. "analyzing reasons for an outcome," "measuring progress," a "Project Appraisal Report," and so on. The recent Panamanian conferences on the state of nutritional field evaluations, reviewed by Henry Riecken, make much the same point: There are different images in our heads about what the process engenders. This variety is not confined to the international sector, judging from recent work for Congress on character of domestic educational program evaluations: We've encountered agency directors who say they don't do evaluations, and the directors' chiefs who say everything they do is evaluation.

The intellective ambiguity engendered by the word evaluation is reflected in language of course. Neologism abounds. We muck about in "formative" evaluation, "impact" evaluation, "social impact assessment," "summative" evaluation, and so on. The promiscuous generation of new phrases in both academic and bureaucratic communities occasionally has positive results: consolidating ideas or marshally energies, for instance.

But this lexical untidiness also has lots of negative effects. There are inappropriate demands for "impact evaluation" of a program in settings where one can't even establish that the program exists, much less a standard for judging impact. Most managers recognize the need to periodically educate their new bosses, new legislators, and their assistants. Assuming some are educable in the time available, explaining evaluation is unnecessarily difficult if one doesn't have a clear idea in one's own mind about what the beast is. There is confusion: a legislature may get a needs assessment from an agency rather than an implementation analysis or audit if its intentions are not explicit. Planning an evaluation budget is a nightmare if one does not have clear definitions in hand, and quality control is absurd without some uniformity. The ambiguity may be warranted in the sense that the history of evaluation is recent and agency staff cannot respond to capricious or ambiguous demands. But it is not prudent if the objective is to routinize the process, to understand its usefulness or uselessness.

For the sake of concreteness, the definition adopted here is that evaluation is an attempt to determine in the least equivocal way possible the effectiveness of an organization's activities in the light of its objectives. The word effectiveness here includes "relevance" and "significance," aspects stressed in USAID's Handbook and Guidelines, in the UN's recent guidelines for internal evaluation (produced by Sohm and Bertra in 1979), in U.S. General Accounting Office work, and in some definitions adopted in other countries, notably by Sang-Chuel Choe of South Korea, Kosit Panpeimras of Thailand, and others. The phrase "least equivocal way possible" reminds us that the evidence ought to be systematic, objective, and interpretable, in plain words. These virtues too are espoused in most handbooks and guidelines.

No definition, accepted widely or not, is sufficient for understanding: One can't build a bridge merely by reading Webster's definition of a bridge. That is one reason for enlarging on the definition a bit to produce a little taxonomy or catalog of types of evaluation.

Types of Evaluation: Needs, Process, Effects

For simplicity's sake, we recognize three functional categories of evaluation here:

- evaluation of characteristics or needs of individuals targeted for a project
- evaluation of project operations: process or formative evaluation
- estimation of project effects: outcome evaluation

Rather more elaborate taxonomies for describing activities often labelled as evaluation have been developed by individuals, by government agencies such as the Canadian Treasury Department and the U.S. Agency for International Development, and by international organizations such as UNESCO (see Soumelis, (1977), for example). Alternatives are discussed after a brief explanation of the types proposed here.

Needs Assessments and Descriptive Surveys

Studies undertaken prior to a project's introduction generally fall under the rubric of <u>needs assessment</u> or <u>descriptive surveys</u>. In simple form, this may be based on the declaration of an expert that a problem exists, and that it is severe. It may include attention to what organizations could do and whether they ought to be doing it. In more systematic form, this may involve the collection of verifiable data in formal surveys, rather than individual opinion. Case studies of a clinical or administrative type may be adjoined to the survey to enrich statistical data. So, for example,

> New Delhi's National Council of Educational Research and Training recommended that systematic surveys of local employment needs be routinely undertaken by colleges <u>before</u> the colleges created new vocational education courses. Because vocationalization is a matter of national policy, and because the policy's effectiveness depends ultimately on community needs, recent NCERT studies of the process of instituting vocational courses in Karnataka focused partly on whether colleges in that area had actually obtained reliable information on employment in vocations before they instituted courses. The NCERT's investigation concluded that insufficient time prevented college's executing good surveys of local need. It is perhaps on account of the failure to do such market research that some vocational programs are of remarkably dubious value. This includes, for example, the invention of courses particularly "suited to the genius of women," such as camera repair. The demand for camera repairers in a country with relatively few cameras and no major production is unlikely to be considerable. (Vocationalization of Education Unit, NCERT, 1978).

Health services researchers in Canada have undertaken limited scientifically designed household surveys to assay the public's acceptance of nurse practitioners, a new method of delivering health services relative to the conventional mode, physicians. The main survey was limited to a large rural area and conducted prior to development of nurse practitioner training programs and prior to later experimental testing of the programs. The survey was designed to establish whether families would accept well trained nurses as a surrogate for physician care, the conditions of acceptability, and conditions under which physicians would be prepared (Chenoy, Pharm, Spitzer, and Anderson, 1973).

Part of the evolution of monitoring and recording systems in South Korea include the recognition that early in the New Community Movement: villagers were sometimes "obliged to carry out projects which were not well suited to their particular situation. Partly on account of complaints about discrepancy between needs and services, there is now an elaborate system for monitoring needs rather closely. A rather long tradition of record keeping and centralized government probably facilitated creation of this system: routine ratings of the state of development of villages in regard to irrigation, bridges and feeder roads, electrification, farm, fishery, and cattle production and support facilities, civil structures and housing characteristics (kitchen, walls, etc.), civil government. The surveys focus on village condition and by implication, needs, are undertaken annually by some 500 teams, and have been running for about six years. That the system is imperfect and may not be applicable in other countries is beside the point for the moment. It does illustrate one approach to understanding character and scope of development problems.

The justification for descriptive information has been reiterated in USAID guidelines such as Turner's (1976, p. 2) and in most good monographs on evaluation. Regardless of the survey's quality, this form of evaluation usually cannot be relied on to show that a new program will work. It is not generally designed to do so. However, preliminary surveys of this sort are frequently undertaken prior to larger outcome evaluation, to understand the nature of the population to which the program is directed, to characterize the scope and severity of a social problem, to identify individuals who may be eligible for assistance from a prototype program, and to obtain baseline data for subsequent attempts to estimate the effects of intervention.

Process Evaluation, Formative Evaluation, Implementation Analysis

The critical examination of the operation of new programs or projects is often labelled formative evaluation, process evaluation, implementation analysis and the like. In simple form, this involves conscientious description of whether a program's operation matches plans, a service, or a product is delivered, and how delivery might be improved. In more compulsive form, it is the "task analysis" of the industrial or human engineer. Where projects or programs are well established, the systematic observation of operations usually falls under the rubric of program monitoring. Where the program is very complex but underlying causal processes are well

understood and measurement systems are very sophisticated, the more elaborate forms of process evaluation are akin to operations research. So, for example:

> The Government of India's recent emphasis on vocationalization has led to the use of colleges in Karnataka as a test bed for understanding problems in the execution of policy. A study to assay problems in implementing policy within colleges alternative approaches to solution has been mounted by the Vocationalization of Education Research and Training. That study involved site visits to all colleges, questionnaire, survey and interview with students, teachers, and administrators, and focused on financial, organizational, and staff adequacy, attitudes of participants, nature of justification for new vocational courses, and the characteristics of students entering such courses (Vocationalization of Education Unit, NCERT, 1978).

> A large class of implementation evaluations consist of the administrative history, a project, reconstructed from interviews, and fragmentary records. The approach is reflected in UNICEF-ICED case studies of Upper Volta's Rural Education Centers (REC). Executed 10 years after REC's inauguration, the studies covered aims of these special low cost training programs for adolescents (literacy and vocation education), calibre of personnel, their training, and tenure, facilities, and daily activities of students in the centers. The cases appear to have been especially useful for synthesizing growth of the centers over a 10 year period in quantitative and qualitative form (759 centers established, financed by Upper Volta, FAO, and others), and major chronic problems. The latter include low salary and sometimes status of center teachers, small size of fields provided by villages for training and self-sust of the center, absence of equipment, and others (Ahmed, Coombs, with Grabe, 1975).

Implementation issues emerge in other areas of course. It is sensible to establish that rural roads which were supposed to have been constructed by local contractors actually were constructed, that the roads still exist physically, and so on. It is similarly reasonable to establish that irrigation ditches and pumps are in place, are functional, and so on, regardless of the durable effects of the installation. Some attention to implementation is essential in nutritional programs to assure that supplements are ingested rather than sold (as in recent Columbian field tests), that contraceptive devices are delivered rather than remaining boxed on a pier (as in early fertility programs in Korea) and so on.

Just as needs assessment surveys are not designed to produce evidence about actual program effects, but may nevertheless be a part of outcome evaluation, process evaluation often provides only a little evidence about whether a project has notable effects. Providing such evidence is usually <u>not</u> the principal function of this type of evaluation and, in any case, the task of determining that some outcome was indeed caused by a specific project requires rather a different approach. This functional distinction

is crucial and has been reiterated in evaluation guidelines of U.S. AID (Turner, 1976, p. 1-2) and USGAO, as well as current evaluation texts. Process Evaluation is also normally warranted, even in crude form, as part of a larger outcome evaluation. In particular, the evaluator ought to have evidence that a program, which is said to exist, does indeed exist and conforms to plans for its operation, if attempts to estimate program effects are to make any sense.

The idea of process evaluation is not new despite the novel labels for the activity. Sensible managers, program developers, or their staff usually trouble-shoot new projects for which they are responsible. What is novel about the enterprise is that it is supposed to be explicit, verifiable, and should contribute to a larger body of knowledge of the project or program. The more typical approach is akin to oral history of primitive tribes: the information is largely in individual's heads, not shared, not especially explicit.

Process evaluation focuses on the match between plan and activity, especially delivery of some good or service, irrespective of the effect of delivery on recipients. But the bounds of this category are blurry. The recent UN guidelines (Sohm and Bertrand, 1979), for example, exclude fiscal auditing and management from their general definition of evaluation, though one might regard auditing as an effort to understand the match between planned budget and actual expenditures. Routine monitoring systems are also excluded. On the other hand, the U.S. General Accounting Office appears to regard auditing as a legitimate aspect of evaluation in principle, as do some operating agencies, though auditing may not have priority in a particular evaluation.

Estimating Effects: Outcome Evaluation

The type of evaluation which has attracted most recent formal interest of government focuses on estimating program or project effects. The phrases usually invoked to describe effects on immediate target groups include summative evaluation, outcome evaluation, and "evaluation of results" (the latter from Sohm and Bertrand). They cover attempts, for example, to design and execute studies of the direct effect of literacy projects on the children to whom the project is directly addressed. The attempt to estimate more distant effects, usually indirect, are often labelled as impact evaluations, as in AID's Guidelines which focus partly on project goals beyond the project's immediate purpose.

Assuring that the project or program is implemented as it is advertised to have been implemented, i.e. process evaluation or implementation analysis, is usually a precursor to outcome evaluation, since estimates of effects will not make much sense unless one knows the program is in place. Understanding the theory or rationale by which a project is supposed to exercise its influence is necessary since this is one major influence on quality of the experimental design.

Of all types of evaluation, this is most often the most difficult: It is rarely easy to estimate effects unambiguously. That meaning well is not enough is clear from efforts to evaluate medical innovations, such as thalidomide, laetrile, and lobotomies during this century, and ptosis during the last. The same humaneness, interest in ameliorating social problems, unimpeded by humane attention to quality of evidence is clear in the invention of worthless education programs and therapies for the mentally ill. There is no reason to presume that most individuals interested in development programs are any wiser than the members of these communities. And the problem of estimating program effects in developing countries appears to be no less difficult than outcome evaluation in other arenas.

To provide estimates which are as unequivocal as possible requires formal attention to the idea of evaluability and evaluation design. Evaluability here refers to the fact that it is simply impossible to come up with sensible estimates of the effects of some programs because nobody's quite sure how the program is supposed to work, much less how it does work, or because no respectable evaluation design was employed when the program was installed to guide collection of evidence. Evaluation design here refers generally to a plan for collecting evidence and assigning goods or services in such a way as to show the magnitude of a program or project effect in the least equivocal way possible, with available resources. Not all evaluation designs are equally good in the sense that the evidence generated by each may be more or less equivocal. Not all designs are feasible in that political, social, or ethical constraints may prevent some from being used.

To illustrate:

The Human Ecology Research Institute's program in nutrition and cultural enrichment was field tested in Cali to estimate the effects of the enrichment/education program on preschool children from the barrios. The test was designed as a randomized experiment in which eligible children were randomly assigned to program or control conditions. The control in this instance is no formal enrichment for at least one year, the program being delayed to this extent for these children. The randomization feature guaranteed that in the absence of a program effect on children's cognitive test scores, the participants and control group members would be equivalent, so helping to assure that the evidence would be interpretable. The experiment was repeated with five waves of children, each tracked over at least five years to assure that evidence on development was unequivocal and relatively precise.

A rural road was constructed between a village and market town to facilitate sale of produce by villagers, school attendance of villagers' children, and public transportation for other purposes. No control villages were found against which to judge the effect of the village. Instead, the standard for judgment was history. That is, the market town's school attendance records for villagers' chil-

dren were sought for a period before the road's installation and for a period afterward. The slight increase in attendance might have been attributable to the road, but seasonal and annual variation in attendance made unambiguous estimates impossible. Fragmentary records on sale of produce, reports of producers and their buyers, and other evidence for a period before and after the road's installation, despite its flaws, suggested that the size of this effect was dramatic and sufficient to override seasonal variation in productivity and sales.

The stress on estimating effects in the least equivocal way possible is a relatively new one for formal government policy. Over the past 15 years, however, the notion has found its way into evaluation guidelines issued by, among others, the U.S. General Accounting Office, UNESCO, OECD, and USAID. That this spirit does not always prevail, indeed cannot always prevail, is also clear. Indifference ("maybe all this will go away"), incompetence (or, more kindly, justifiable ignorance), and corruption (fraudulent or merely notably less than candid) account for many poor evaluations during the 1960's in the U.S. We have every reason to expect that the same difficulties will emerge in other contexts.

Other Ways of Specifying "Types of Evaluation"

The three relatively simple forms of evaluation underly more complex forms. One may enlarge on the effort to estimate program or project effects by obtaining costs as well as benefits, and exercise a bit more arithmetic skill to compute a cost/benefit ratio. Naturally, one type of evaluation does not preclude another. And most good general evaluations combine all three types in an effort to understand who needs the service, whether the service was delivered, and what the effect of delivery is.

Moreover, any of these types can be mapped onto other taxonomies or categories for describing evaluation activity. The USAID Handbook, for instance, recognizes project level, sector level, and country level evaluation. The same structure is espoused by, among others, Kosit Panpiemras in describing Thai approaches in recent UNESCO conferences on evaluation techniques. One may then mount needs assessment at local, regional, or country level. This geographic taxonomy is helpful to policy-makers who must deal with all three levels at least insofar as it implies level of difficulty and complexity of an evaluation. Control over quality of the evaluation design, the quality of records and other evidence, links between cause and effect, and so forth are rather more difficult as one gets further from the point of activity.

An alternative to the geographic taxonomy of types of evaluation, and perhaps a more useful one, hinges on users of evaluation results. This framework specifies that regardless of the type of information one obtains, the results are classifiable into those which are useful to: the policy-maker, to the administrator-manager; to an oversight or audit agency; to the scientific/developer community; and to public interest

groups. So, for instance, the local public may be interested in better description of local needs; at the policy level the specific locale's needs are less interesting than needs in the aggregate. The manager may be considerably more interested in process evaluation or trouble-shooting because it holds the promise of improving the program, than in outcome or summative evaluation. This user-oriented taxonomy is related to draft proposals developed by the U.S. Department of Health, Education, and Welfare, by staff of the U.S. General Accounting Office, and others. It is less useful perhaps in other countries, such as Pakistan, where access to evaluative information is not unconstrained, implying only one central user group.

Malaysia's approach to classifying evaluations is based more on temporal characteristics, than on function or type of information collected. Mohr Abdul Ghani's description merely specifies ex ante (planning feasibility) assessments, in vivo evaluations (monitoring and trouble-shooting), and ex-post facto (after the fact evaluation). The staff of his Socio-economic Research and Planning Unit apparently find this useful in both planning and monitoring evaluation activities.

The point is that other taxonomies of "types of evaluation" are possible. Most variations other than the one suggested here depend on a particular government structure. It is partly on account of this variability across countries, and in the interest of common language of a durable sort, that the three elementary types have been discussed here. They are simple, sensible, and can be dovetailed easily with other taxonomies.

Distinctions and Lack Thereof

The distinction between planning and evaluation is often made, though the activities implied by each will be identical in some instances. To the extent that planning is defined solely as "how we expect to deliver program services," the overlap is insubstantial. To the extent that planning is defined so as to include pilot tests, needs assessment surveys, evaluation design and so on, planning coincides with evaluation, and the distinction is meaningless.

Whether evaluation is a management tool or research tool is a frequent topic of discussion, e.g. at UNESCO's Kuala Lumpur conference on evaluation techniques in Asia. The argument often takes the form of distinctions between "decision-oriented" work, which presumably managers do, and "conclusion-oriented" work, which presumably researchers do. This way of framing the issue is as common within the U.S. as it is elsewhere, if we may judge by statements of the U.S. General Accounting Office and the U.S. Office of Management and Budget, on the role of evaluation in that government's policy. The issue seems to emerge partly because we often need to link unfamiliar terms or activities, such as "evaluation" with more familiar ones, such as management. The issue is gratuitous to the extent that there is no reason for decisions to preclude understanding, or conclusions preclude decisions. It does make sense to regard the relative emphasis

on decisions and conclusions as one of degree, and to regard evaluation as both a research and management enterprise simply because it is. The implications of this distinction, whether one thinks it's trivial or not, are important. For example, budget cuts are normally made to research; calling evaluation a management tool helps to avoid an agency's having its budget cut. On the other hand, calling evaluation a management tool raises the formidable spectre of management or parliamentary decisions made on the basis of evaluation, and this may engender fear and corruption. One may then argue that evaluation ought to be classified as a research enterprise, on which no decisions are based, so as to assure candor and cooperation in obtaining good evidence on program effects. This issue is not yet resolved.

2. Introduction to the Logic of Evaluation Designs for Estimating Program Effects

This section concerns the logic rather than tecnical aspects of evaluation design. By "evaluation design" here, I mean a plan for structuring data collection and assignment of services to individuals or regions so as to produce the least equivocal evidence possible.

There are several reasons why the topic, despite its heavily technical character, the press of program-related problems, and the immediate demands for ex-post facto evaluations, ought to be understood by managers and others whose responsibilities are not technical.

First, the ambiguity of evidence hinges on design. And insofar as the manager is responsible for producing interpreting evidence, he or she has some responsibility for evaluation design. As a practical matter this often means providing logistical support to the evaluator, and more important, developing mechanisms which provide services in accord with the evaluation design and with the program design.

Second, a fair number of very common evaluation designs yield estimates of program effect which are very likely to be biased negatively. That is, programs which in fact have little or not effect appear to be harmful; programs with small effects appear to be worthless. It does seem to be in the manager's interest to know something of these problems, if only to avoid embarrassment. It is clearly in the evaluator's interest to know about them.

Third, the logic of evaluation design for social programs is a rather remarkable by-product of methods developed in this century to understand the effectiveness of new surgery, drugs, fertilizers, feeding schemes, and other physical phenomena. The problems of assuring good evidence in these are not much different in character from those encountered in the social sector, though there is clearly a difference in difficulty of the problem. And the manager who is sensitive to history and philosophy of science may wish to understand the ideas.

Finally, the evaluation designs described here are understood by agencies which oversee the quality of outcome evaluations. This includes the U.S. General Accounting Office and its consultants, and special oversight divisions within AID, UNESCO, and other agencies. It is unlikely that ignorance of the logic is in the manager or evaluator's interest.

The following remarks cover major designs, together with notes on applications and variations.

Design of Outcome Evaluations

The state of the art in evaluation designs was well articulated in a variety of recent texts, monographs, and applications are reported in contemporary journals such as Evaluation Quarterly. Consequently, treatment of the topic here is brief and focuses on the standard of reference for estimating program effects. The integrity of any outcome evaluation hinges on the standard of comparison used as a basis for estimating program effects. To the extent that the choice or construction of the standard is poor, the results of evaluation will be ambiguous or misleading. Development of a standard lies at the core of contemporary work on evaluation design.

The closing paragraphs of this section provides references to recent work on managerial, political, institutional, and scientific problems encountered in evaluations, and on solutions to them. The case studies presented elsewhere also illustrate some of these matters.

Randomized Experiments

In randomized experiments, some individuals (or regions, or institutions) are accorded services through a new pilot project and the remainder receive conventional services. The latter group, receiving services through ordinary means, constitutes the standard: a control group against which the new project is judged. More important, the eligible individuals are randomly assigned to one or the other program so as to assure the equivalence of the two groups in the long run. The equivalence produced by randomization decreases the ambiguity of subsequent results.

For instance, half the children in an eligible group might be assigned randomly to an instructional program on (say) hygiene using radio broadcasts and the remainder to TV based instruction on the same topic. The main purpose of the randomized assignment is to assure that, in the long run, the children in each group are equivalent, aside from the differences induced in their performance by the different modes of instruction. When properly executed, the design furnishes an estimate of relative effectiveness of the two modes of instruction. The estimate is relatively unequivocal in the sense that differences in test performance between two groups reflects differences in effectiveness in the instructional modes rather than to preexisting or natural differences in the groups, differences in the reliability of measuring achievement in the two groups, and so on. The latter are in principle removed by the randomization process.

The design has been used in a remarkable variety of settings.

In Mexico City, for example, several schools have been involved in estimating the effects of Plaza Sesemo, an educational TV program for pre-school children, on the cognitive abilities of the children. Children were randomly assigned to classes in which they watch the program and to a control gruop in which no viewing is involved. This is analogous to experiments in some cities of the United States where families were randomly assigned to a

condition in which they were strongly encouraged to have their children view the program or to a control condition in which they were not encouraged.

Early efforts to determine whether cloud seeding with silver iodide crystals or other substances affects rainfall were based on evaluation designs which involved analysis of time series data on rainfall, and search for increases in rainfall following seeding. European and Asian researchers found that the effects of cloud seeding programs could not be estimated unambiguously using the approach largely on account of complex weather cycles which could not be adjusted for in analysis. More recent evaluations in Europe depend on randomized experiments. Of all short periods (each a day long, say) predicted as being "suitable" for seeding some are randomly assigned for cloud seeding, and the remainder serve as control periods. Estimates of program effect are based on comparison of hail and rainfall following seeding periods against hail and rainfall following control periods.

Eliciting candid information about fertility control practices from women is difficult in many surveys. In the simplest cases, women may be embarrassed about responding to an interview, or cultural and religious mores may foster misleading response. To get round the problem, so-called randomized response methods have been field tested in the U.S., Taiwan, Canada, and elsewhere. A simple variation on the technique involves first presenting the question, e.g. Did you use this contraceptive device this year? The respondent is asked not to respond directly but rather to roll a die. If the die turns up 1, the respondent must respond truthfully with a Yes or No. If 2, 3...6 show, the respondent answers Yes. The interviewer is not shown the die, nor told anything about the true state of the individual. With a large sample and adherence to the regimen, it takes only simple probability theory to estimate the true proportion using the contraceptive, i.e., the proportion of false positives is known and adjusted for in estimation. Most good tests of the method involve randomly assigning respondents to either this new method or to conventional interview. Estimates of the effectiveness of the technique are based on comparison of each group: estimated contraceptive use, attitudes of respondents toward the method, and so on.

Randomized experiments are not always appropriate or feasible. Appropriateness depends on whether anyone is actually interested in estimating a program's effect well, and how much that information is worth in the long run. Feasibility depends on competence of the evaluator and of the manager responsible for program and evaluation, on control over the randomization process, and on political-institutional factors. The latter includes resistance to comparing a program group to a "no program" control group. The resistance has been degraded by setting up comparisons of program variation A to variation B, where variations differ in cost, by evaluating expensive

components of programs rather than the entire program, and other strategies. The approach is most often feasible when services or goods are in very short supply, the effect of services is not known, and equity invites random allocation of the service or good to equally eligible individuals.

Time Series Designs

In some program areas, considerable data on the outcome of interest exists. So, for example, some cities have maintained very accurate monthly records on the incidence of mortality and certain diseases for 10 or more years. The existence of such records makes the possibility of a time series approach to estimating program effects possible.

In particular, inhabitants of a target area are monitored at successive points in time, the program is introduced, and the area is again monitored. The estimate of program effect is normally the difference between the actual outcome level, of mortality, for instance, following the program's introduction and the level projected from prior years. The standard of comparison in this case is historical: a prediction, based on prior data, about what mortality level would have been without the new program. Such historically based comparisons have the benefit of economy, and to the extent that records are uniformly accurate, and the time series is free from peculiar and poorly understood seasonal variation, the estimate of effect will be good. To the extent that quality of records varies, that factors other than the program exercise an unknown or incalcuable influence on mortality, or that cyclic variations are not well understood, the estimates of program effect will be ambiguous at best and misleading at worst.

So for example,

To determine the effectiveness of more stringent laws governing speeding and drinking while driving, for example, in some of the United States, times series data on traffic accidents, arrests, and the like have been used. Variations in the quality of crime records and temporal variations in definition of crime make it more difficult to use archive records for determining the effects of new law on crime rate. But some attempts have been promising e.g., estimating the effect of more severe sentences on carrying an unlicensed pistol in the state of Massachusetts.

Analogous approaches to estimating the effect of new law have been undertaken in Denmark. For example, the effect of relaxing pornography laws in Denmark was established in part by examining time series data on sex-related crime rates. The Kutchinsky studies of the matter are especially conscientious because he recognizes that variation in legal definitions of crimes, variations in cultural attitudes and propensity to report crime. He documents the extent to which these are plausible influences on the outcomes.

The two approaches, randomized experiments and time series are not inimicable, of course They can be combined in a single evaluation and often are. The main point is that the randomized experiment sets up a contemporary empirical standard for judging the effect of a program. Effects are based on comparing a group of people who participated in program A against an equivalent group who participated in B. The time series, on the other hand, involves a historical standard and assumption which is sometimes less tenable: that the behavior of the area or group under examination can be predicted confidently from earlier behavior.

Most readers will recognize far better than I that the time series approach is feasible in developing countries only when the pertinent records are available and this is exceptional. Some Indian cities have fairly good records of mortality and birth rate, but quality and availability vary considerably over the country. This makes local evaluations possible occasionally, but national evaluations very risky. Similarly, though Malaysia's Decision Support System is very elaborate and involves some controls over reporting quality, the data are still too recent to use in time series anlayses. There are simply too few time points to exploit the statistical machinery appropriately. South Korea's system for reporting from local levels, a longer tradition of reporting makes some time series assessments possible in health, for example, but the data are insufficient for others.

Designs based on Allocation in strict Accord with Measures of Need or Merit: Regression-Discontinuity

The available archived data are often not sufficient to employ a time series approach. And, though opportunities for executing randomized field tests exist, the strategy is not always feasible. In those cases where people, groups, villages, or regions are assigned to receive services in strict accord with measurable need, however, a reasonable alternative design can be exploited. The design requires measurement at least two points in time, on a sample of both service recipients and non-recipients and works in the followng way.

The need of regions or individuals is first measured, e.g., health status is measured for each of 50 sectors of a country. The neediest sectors are assigned to receive program services strictly on the basis of their need. For example, only the three with the highest mortality rates receive services. After the service program has stabilized, all sectors are measured again. The actual estimate of the health program's effect hinges on the assumption that there is normally a strong relation between health status from one year to the next. To the extent that the program disrupts that relation, e.g., influences mortality rates of the selected areas notably, the program can be said to have been effective. The estimate of program effect is then judged relative to a projection based on one prior measure. Figure 1 illustrates the null condition, i.e., the presumption or standard against which program effects will be judged. If the program's introduction engenders a disruption of the relation, then Figure 2 will appear.

To illustrate:

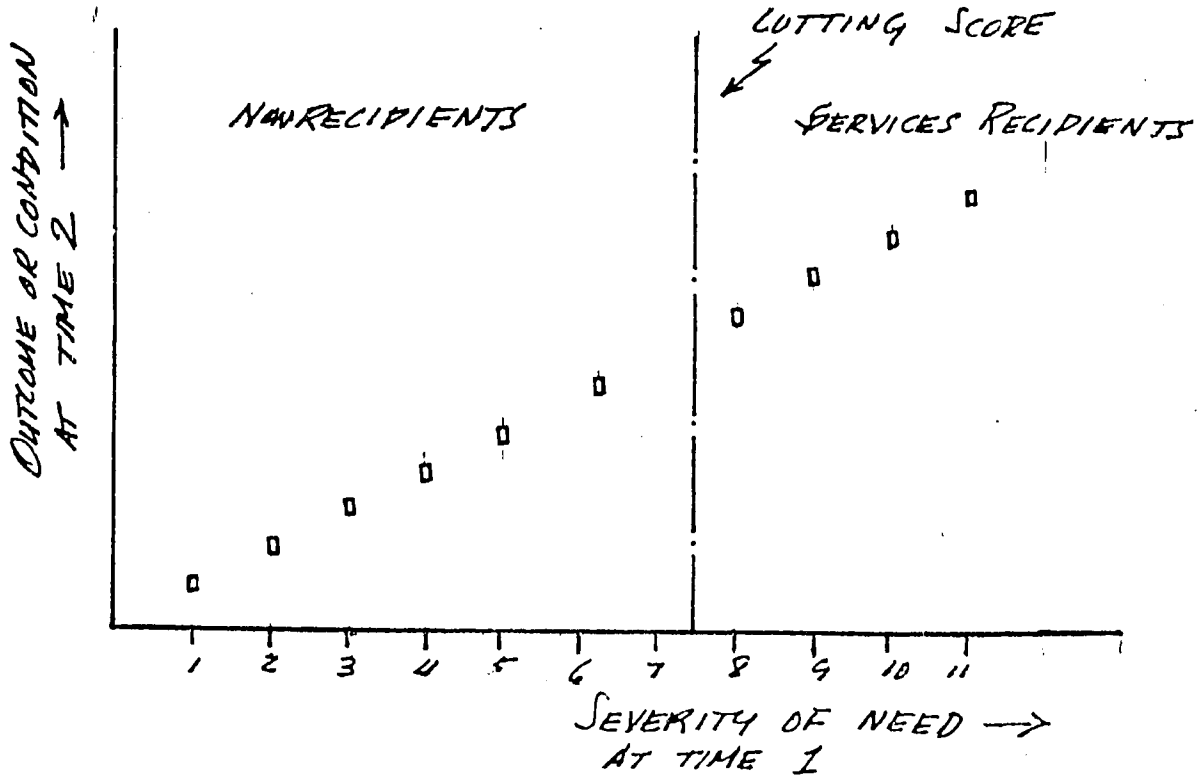

FIGURE 1. NULL CASE: NO PROGRAM EFFECT

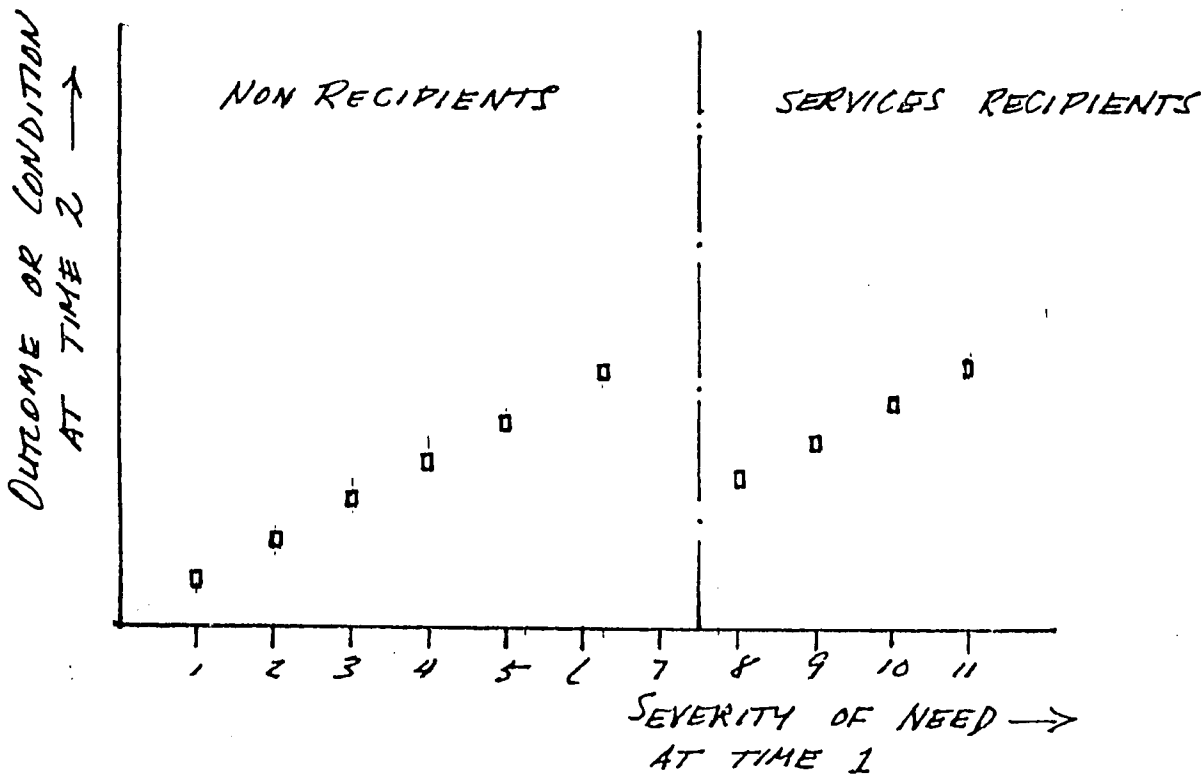

FIGURE 2. DEMONSTABLE PROGRAM EFFECT

The support of rural road construction in developing countries is often undertaken partly in the interest of improving conditions of the owners of small farms. In planned evaluation of the effect of construction then, one might partition a country into geographic sectors characterized by predominance of small farms. The income, productivity, etc. of such farms are in principle measurable and these indicators, taken alone or in combination, may be used to assign priority. Road construction resources are then assigned in accord with priority, aside from predictable political exceptions and under logistical constraints which are specifiable beforehand. Later measures of income, say, are then obtained on a sample of areas in which roads have been constructed and areas which fell below the point of clear need. The relation of before measures to the after measures will be discontinuous if the roads had a detectable effect on income and its magnitude should be estimable.

Use of this design is not yet common, but it has substantial merit on political and technical grounds. If the relationship of needs in one period to the next is strong and the program is assigned strictly in accord with need, the design is at least as good, on technical grounds, as a randomized experiment.

Before-After Designs with and without Arbitrary Comparison Groups

The design for randomized experiment provides a guarantee that if the assignment is indeed random, the groups so formed will be equivalent in the long run and in the short run any differences which do appear will be due to chance alone. The time-series and regression discontinuity approach require rather more tenuous assumptions, notably that prior behavior is a very accurate indicator of what would happen in the absence of the new program. Those assumptions are tenable to the extent that the data are adequate and prior conditions well understood. In some setting, none of these approaches will be feasible, and instead weaker evaluation designs, requiring more assumptions, which may or may not be verifiable might be considered.

One of the simplest of such evaluation designs involves measurement of conditions before and after a program's introduction. The simple difference between the two is typically used to index the size of the program's effect. The design is very risky in that results may be, and often is, misleading.

For example, in attempting to assay the effects of fertility control programs in Korea, evaluators used simple change in rate over five years to index the program's effect. The difference between measures was not substantial and indeed was disappointing. The disappointment may not have been warranted in that the preprogram measure could have been the product of a long trend upwards in fertility rate and the program helped considerably to deflect the trend. Or, even more disappointment may have been warranted if the preprogram

measure was the product of an earlier downward trend and the other events coinciding with the program actually induced higher fertility rate. The first explanation is more plausible. But the point is that pluasible competing explanations exist and cannot be accommodated using the before-after data alone.

Dessaint, et al (1979) summarize what appears to have been a fine before-after study of the effect of rural road construction in Madagascar. In 1964, a French group undertook an elaborate socio-economic survey of the Andapa Basin in exploratory research. In 1970, the World Bank initiated a rural roads project in the area. The prior data was discovered after the fact, and its availability led to surveys in 1974-75. The before-after comparison provided informative and, by and large interpretable, estimates of the effect of road installation at the regional level. Both the World Bank report and the Dessaint et al review stress that the problem of no baseline data is chronic and often leads to idiosyncratic and unverifiable judgments about the effect of road construction. Even with the before-after measures, the possibility of cyclic variations in regional economy, rainfall, etc., and the absence of control areas makes inferences more tenuous than they might otherwise be. Further, external factors such as country economy or world market for region's products may bury the effect of roads entirely.

There do appear to be special settings in which the problem of competing explanations in before-after studies are less severe. Consider, for example, technical instruction. It's reasonable to suppose, that a notable change in the test performance of a large group of young children, which occurs after the introduction of a new fluid mechanics irrigation course in their classes, is attributable to the new program since most children do not have access to such training materials and do not usually learn the technical aspects on their own. Determining whether courses in welding, in watch repair or plumbing, and so on have their intended effect on students' knowledge can often be evaluated along the same lines. The key point is that competing explanations for an increase in ability are negligible in some settings. This makes before-after design less objectionable.

More often, it is reasonable to create a comparison group, in addition to using the before-after approach. The comparison group becomes a complementary device for predicting what individuals would have done if they didn't receive program services. The utility of this standard, of course, depends on comparability of the comparison group. Randomized experiments generate groups which are equivalent within specifiable statistical limits. Choosing a comparison group after the fact does not provide any similar guarantees. In particular, the utility of comparison groups so constructed or selected depends on knowing that all the factors which influence equivalence on the outcome variables are known. The assumption is major and it is usually difficult to verify despite the commonness of the design. Unverifiable and unrecognized differences between those individuals involved in a new program and those who were not are the rule rather than the exception.

Ex-Post Facto Studies

Ex-post facto studies generally refer to efforts to understand a project's effects when the evaluator played no role in designing the program evaluation prior to program installation. As a practical matter, the phrase often means that the evaluator obtains as much data as possible bearing on program effects during a site visit or similar investigation.

Where the project staff have executed an evaluation design, the task is relatively easy. The evidence, design plan, etc. are merely consolidated by the evaluator and analyzed independently. Where project staff did not design and execute a formal evaluation, then the task of outcome evaluation is considerably more difficult. That is, if no evaluation plan was laid out initially, if no data were collected in accord with the plan, the evaluator must undertake the task. The severe disability under which the evaluator operates in this case is that some evaluation designs, the best ones, cannot be executed after the fact. A randomized experiment cannot be mounted, a regression discontinuity approach usually be exploited, and so on.

This is not to imply that ex-post facto studies are useless. They often have clear merit in establishing that the project is important, relevant to local needs, and has been implemented more or less as advertised. It is to say that estimating the project's effects in the least equivocal way possible requires planning. If this is not done before the program's installation, the ex-post facto study is more likely to yield ambiguous results. If the ambiguity of evidence plays a secondary role, and (say) the expert judgment of the evaluator, though unverifiable, plays a primary role, then the ex-post facto study is more acceptable.

Goal Attainment Scaling

The phrase goal attainment generally implies achieving some specifiable objective. Strictly speaking the rubric implies nothing about the logic of evaluation design. It merely identifies outcomes to which the design must attend. That is, merely making goals explicit and tracking what appears to be progress toward them is sensible only if cause-effect linkages are clear. Most often they are not. In the absence of a formal design to help tease out competing explanations, "achievement" may be a matter of chance, novelty, and simple growth processes independent of the program.

The evaluation designs discussed here for reducing ambiguity of evidence are compatible with a goal emphasis. Time series data, for instance, permit one to estimate what would happen without the program, and any sensible goal setting relies on that projection. Similarly, one might invent method A to teach irrigation farming better than the current method B; a randomized experiment might then be used to assure that A is indeed better.

State of the Art

The ideas of formal design of outcome evaluations and of initiating the process before a new program is introduced are fundamental. After the fact examinations, like autopsies in medicine, can rarely be as informative as planned experiments regardless of their sophistication. More important, the state of the art in design of evaluations has improved dramatically over the past few years. A few monographs treat managerial, institutional, political and other problems engendered by high quality evaluative research. Riecken et al (1974) make a deliberate attempt to recognize such problems and to enumerate solutions to them. Rivlin's (1974) volume is consistent with this one, but is more policy oriented. The stress is on evaluation in the United States but illustrations from other countries are included as well. State of the art volumes such as these update earlier monographs by Suchman (1967) and Chapin (1947). More recent sources of information are readily available. Evaluation Studies Review Annual and Policy Studies Review Annual each reprint a variety of best articles on related topics published in journals or books over the previous year.

Similarly, the methodology is being articulated better. The annual reviews just cited constitute one source of information on contemporary developments. Other volumes are dedicated to special interest areas, e.g., Adams (1975) in evaluating criminal rehabilitation programs. Still others enlarge, conscientiously and in imaginative ways, on the early state of the art monographs produced by Campbell and Stanley, e.g., Cook and Campbell (1976) Caporaso and Roos (1976). The Campbell-Stanley monograph is available in several translations and they are listed in the references.

New professional organizations have been developed to facilitate discussion of problems and alternative solutions. This includes Council for Applied Social Research in North American and Europe, the Evaluation Research Society in the United States and others.

Tailoring Designs

It doesn't take much field experience to recognize that the broad designs discussed here must be fitted to the situation at hand, modified to suit political, logistical, and other constraints. Those aspects of design which might be manipulated to improve feasibility are discussed briefly below. Again, their examination and application require expertise which managers do not have. Specialists are warranted and to the extent that one cannot capitalize on local universities, census bureaus and the like, the ability to tailor designs is lower and the quality of the evaluation will follow suit.

The characteristics of design which may be manipulated include:

- Component-wise evaluations
- Unit of anlaysis
- Time
- Side Studies
- The Program

Each is described briefly below

Componentwise Evaluations

There is no natural law which says that every feature of a program can be evaluted with the same level of confidence. Some components may be evaluable; others may not. Recognizing this is important, enhances sophistication of design, and enlarges feaibility.

So for instance, one may be fairly confident that in evaluating an integrated rural development project, a before-after design to estimate the effects of a reforestation program may be quite adequate. That is, the effects of the reforestation component can be established with some confidence using the design, despite the fact that all other components are not evaluable or their effects cannot be separated from other influences using any design.

Similarly, the Cali, Colombia experiments on nutrition and education were designed so that all malnourished children in particular sectors receive nutritional supplements regularly, and measures of head circumference, weight, and other indicators of effects are measured regularly. The evidence is equivocal to the extent that even malnourished children grow and progress is attributable in part to natural growth. It is less equivocal to the extent that local norms for growth exist and can be used as a standard; but we are less confident to the extent that the norms are outdated, based on groups very different from the children at hand, etc. The education component of the project has been evaluated using a design which will yield less equivocal results: a randomized experiment in which educational services are delayed temporarily in order to verify that the educational innovation does what it was intended to do.

Unit of Assignment and Analysis

In testing new education programs, one may estimate program effect at the individual level, at the classroom level, or at the school level. The implications for design of, say, a randomized field experiment are that one may

- assign individual children within a classroom randomly to each of two alternative education programs to determine which is best, as is often done in the U.S., or

- assign groups of children, where groups are defined by neighborhood for instance, to each of two or more educational regimens to determine effectiveness, as in the case of the Cali experiments, or

- assign schools, randomly to each of two educational regimens to estimate average effectiveness as in the AID supported.

The implications for a regression-discontinuity design are for instance

- employ a regression-discontinuity design for evaluating the average effect of a set of new roads on rural areas, or

- evaluate any given road, using a before-after design, recognizing increased ambiguity.

Time

Most people recognize that time of measurement is critical in most evaluations. The effects of programs may appear small if measurement is premature, e.g., measuring infant mortality a month after a clinic is installed. They will also appear small if program effects decrease over time, e.g., the effect of arithmetic education on third graders.

Because the temporarl character of effects is often unknown, the ideal is to measure as often as possible. Where that is not possible for everyone, one takes a sample of people, or a sample of time. The options here then are all options that apply in conventional statistical technology to sampling problems.

Side Studies

There is some good justification for the idea that side studies are essential in most evaluations, though their product is indirectly related to estimating impact, and for future evaluations.

Consider the simple matter of validity of data on infant death rate. One may rely on government records, but they are often imperfect in developing countries. To the extent that they are imperfect, any evaluation will be less sensitive to program effects than they ought to be. One may then also collect information directly in interviews using indigenous staff, recognizing that this too is imperfect: people do forget that a child died. But the combination of both kinds of information can yield better estimates of mortality if the sources of inaccuracy in each are independent. And that will lead to more sensitive evaluations. Moreover, differences in the two sources of information are important, and have implications for designing better evaluations in the future. This information sharing is critical. The state of the art in combining imperfect sources of information is advancing rapidly, thanks to cooperative census work in Morocco, Libya, Korea, and elsewhere; the strategy falls under the rubric "dual system estimators."

Program Variations

3. Ambiguity and Bias in Estimating Program Effects: The Problem of Competing Explanations

The stress here is on evaluations planned specifically to obtain relatively unequivocal and unbiased estimates of program effect. By "unequivocal," we mean that if one announces, say, that the effect of a fertility control program decreases birth rate by 10%, then the effect ought to be clearly attributable to the program's influence rather than to other extraneous influences on measured birth rate. The latter might include, for example, simple biases in reports by families to an interviewer or census taker. They might include a complex cycle in birth rate in which the low point happens to coincide with the time at which the interviews were undertaken. They might include considerably more complex influences, such as a poor match between a sample of women who've participated in the family planning program and a comparison sample of women selected after the program's introduction, who have not had the interest in or opportunity to participate. Such influences can be disentangled from the program's effect provided that the evaluation design is of high quality and its execution begins before the introduction of the program.

The methodologist's concern about bias in estimates of program effect has similar origins. Merely assuming that a 10% drop in fertility rate is due to a new fertility program is infrequently sufficient. The drop may in fact be attributable to trends, the peculiar composition of the sample, and so on. To the extent that the evaluator fails to recognize these and estimate the magnitude of their influence, the 10% figure will be a biased estimate of the reduction in fertility rate.

Quality in evaluation design, indeed the fact that evaluations must be designed before programs are introduced is fundamental to obtaining reasonably unequivocal estimates on program effects. To the extent that the evaluation design is poor, simplistic, or absent, then subsequent analysis will be difficult, and conclusions will be misleading at worst. Projects may be judged worse or better than they actually are: The theme is emphasized in most guidelines on evaluation, e.g., Turner (1976, p. 3), U.S. AID (1976), USGAO (1978), and contemporary texts.

No outcome evaluation of a social program is completely unequivocal. That is the reason for using the word "reasonably" unequivocal in these remarks. The evaluator's objective is normally to reduce the uncertainty to its lowest possible level available with resources.

Much of what has already been said about evaluation design, of course, concerns data analysis in one way or another. The following discussion is tied less to design than to problem of making sense of the data stemming from nonrandomized studies, including surveys. The review is brief but hopefully not provincial. It ranges from the simpler devices, such as checklist approaches to making inferences about a program's impact, to those statistical methods which yield a better-qualified inference but are technically demanding. None guarantee unbiased estimates of program effect nor do they guarantee that competing explanations will be eliminated. But all help to establish the plausibility of any given interpretation of the data.

Checklist Approaches

Given some data bearing on a program's impact, the idea behind many checklist approaches is to make some judgment about the existence of an effect, then to enumerate all the competing causes or explanations for the effect other than the program itself. The final step is usually to attach some level of plausibility or import to those competing explanations.

Consider, for example, data on a novel treatment for schizophrenics undertaken in one ward of a hospital and data on schizophrenics treated in another, nonequivalent ward. One might first look at levels of severity following treatment for each group and infer that because severity is higher for the novel treatment group, that treatment harmed rather than helped people. There are a number of competing explanations for a between-group difference. of course. The type of patient normally assigned to the novel ward may, for example, differ from those assigned to the regular ward. And enumerating these explanations is easier if we have at our disposal a checklist of "threats to internal validity" such as the ones enumerated by Campbell (1957), Campbell and Stanley (1966), and Cook and Campbell (1976). Their lists are elaborate and well articulated elsewhere. so only a brief description of major themes is provided here.

The maturation of each group may differ notably. Participants in the novel program, for example, may have been deteriorating over the past few years while the comparison group condition may have been relatively stable. The selection process for each ward may be deliberately or accidentally designed to assign those with differing severity to each ward. Even when their history is similar, if initial severity is different in each ward, then their rate of deterioration may differ irrespective of the program's impact. Testing patients in the new program may also produce artificial effects e.g., by heightening patient and staff awareness of their condition. And if the instrumentation changes during the process, if different raters are used at different stages of the evaluation, then the resulting differences in ratings may be mistaken for program effects. So-called regression artifacts may also be critical: If individuals are assigned to the new ward on the basis of their high initial severity, in the absence of any program effect and any major trend in the overall population, that subgroup will have a lower severity score at the second testing. The drop is attributable to a regression to the mean which affects any measure with imperfect reliability, rather than to the program. Dropout rate or attrition may produce artificial differences between groups if, for example, the least severely ill members of the new treatment group reject the new treatment and leave the ward and perhaps even check themselves into the regular ward.

The recent list of threats to validity issued by Cook and Campbell (1976) is complex, and its completion can be very tedious. Nonetheless, in the informal impact evaluation, an approach like this can be enormously helpful in avoiding egregious errors in literal or quantitative analysis of the data.

Approaches Based on Stereotypical Errors

Any checklist is inefficient to the extent that the list is long and contains items which are irrelevant for the evaluation at hand. To accommodate the problem, it is sometimes possible to identify stereotypical ways in which inferences about program impact may be wrong and so truncate the list to manageable size. A given set of literal stereotypes may apply to only one class of evaluative settings, e.g., in compensatory education, and irrelevant to another, such as irrigation. But they can be more useful in developing prescriptions which are quantitative and qualitative, rather than merely qualitative, to deal with the problem of estimating impact.

The basic objective is to recognize chronic problems of data analysis in a given substantive area and to determine their severity for the evaluation at hand. For example, Campbell has identified six such problems, and Campbell and Boruch (1975) furnish some procedures for detecting and accommodating them. Those problems identified as chronic in the evaluation of compensatory education programs using covariance or matching designs and observational data, are generalizable to other settings. They include regression effects, differential growth rates, differences in reliability of measurement over time and with subgroup, and ceiling and floor effects on measurement.

To make matters concrete, consider an actual cultural enrichment program to which economically deprived preschool children have been exposed. The comparison group, sleected after the fact, consists of children from families whose economic condition is better and who are thus ineligible for the program. The specific mathematical character of the setting, given in Campbell and Boruch (1975), is essential to estimating program effects but not essential for understanding the problems and prescriptions which we give next.

The regression-effects problem turns on the fact that unreliable measurement in surveys induces biases in estimation of program effects when covariance, regression, matching, or similar conventional techniques are used. So, for example, a covariance technique or matching might be used by the conventional analyst to adjust out pre-existing differences between a treated group and the control group, without recognizing that imperfections in measurement of the covariates or matching variables often lead to estimates of program effect which are seriously biased. Some conditions, which are common but do not always prevail, can lead to estimates of program effect which are biased downward i.e., estimated effects are small even where actual effects are large, and estimates are negative, making the program look harmful when actual progarm effects are negligible. The conditions under which the problem occurs are described mathematically and literally, together with examples, in Campbell and Boruch (1975).

To help assure that the problem is detectable, it is essential that the program evaluator acquire information on (a) the process underlying assignment of program and comparison-group members, (b) statistical data bearing on the character of both groups before the program begins, and (c) reliability of matching variables or covariates. Estimates of reliability can sometimes be used to adjust conventional analyses for the presence of measurement error. The other information is helpful in understanding whether particular approaches can or should be used by the analyst.

Differential growth of the treated and comparison groups appears to be a chronic problem which cannot be remedied without strong theory and good data on growth processes. When tested repeatedly over time, for example, middle-class children's achievement test scores often increase more rapidly than do scores of children from economically deprived settings. The stereotype occasionally identified here is a "fan spread" pattern of average or mean achievement for economic groups when plotted over time. If the phenomenon goes unrecognized by the conventional analyst, then estimates of program effect will be biased. An analysis which looks only at change scores in the program group relative to the middle-class children will lead to a declaration that because the middle-class group is increasing at a faster rate, the program must have had a negative impact on participants. The declaration will be wrong unless the differential growth rate has been recognized and accommodated.

There is some evidence to suggest that the less privileged child is measured with somewhat less reliability, using standard achievement tests, than the middle-class child, and further that the reliability of measurement of each group will (depending on age, grade, and other factors) increase as children and testors become more familiar with the test. One of the stereotypical problems which may crop up in the field setting then is that the differential rate of errors in measurement produces a small difference between groups when in fact the difference is large. That is, the unreliability obscures real differences, especially on initial testing. If children are tested a second time and reliability of measurement of both groups is higher on that testing, the difference between groups will, under fairly common conditions, increase. The increase will be attributable to the properties of the test and to practice effects rather than to any real program effect. But if the differentials in reliability of measurement go unrecognized by the conventional analyst, then the post-program differences may be declared (erroneously) as indicating that the program had a negative effect. Again, the exact conditions under which the problem occurs and evidence are given in Campbell and Boruch (1975). The simplest prescription that we can offer to establish the existence and magnitude of the problem is to obtain hard data as a check on its character: pilot tests or side studies in which intergroup differences in reliability or intertime differences are estimated, on-site studies of the test process, post-facto estimates of the internal consistency of test responses estimated by group and by time period. Without the estimates, it is unlikely that adequate corrections can be made to conventional analyses.

Floor effects refer here to the problem of inadequate measurement of children whose ability level is considerably lower than that measured accurately by the achievement test used in the evaluation. That is, children in this range score near the zero or chance level on a test not because they know nothing but because the test is insensitive to ability in the lower range. The consequences can be obvious: Even if the compensatory program has a notable effect, that effect may still be too small to be reflected by an insensitive test i.e., the effect may not be sufficient to boost children into a measurable range. The problem is not a new one, nor is it confined to education. It is a dangerous one to the extent that conventional analysts

who do not recognize it will be tempted to declare that a program estimate which is near zero means the program had no effect, rather than that the measurement was faulty. Worse, the problem can lead, under certain conditions, to estimates of effect which are negative. Again, the only reasonable prescription to follow, especially when the tests' properties are not well documented or are suspect, is to run pilot tests prior to the program evaluation to establish their sensitivity and validity.

To summarize, it is clear from case studies of evaluations that some conventional data analyses are susceptible to stereotypical problems which often go unrecognized. Those problems can be anticipated and documented with side studies and pilot tests. The idea to identify stereotypes, can be helpful in avoiding serious problems in estimating program effects, especially when unrecognized problems engender estimates which are negatively biased. The hazard, as in any such effort, is that the stereotypes themselves may be irrelevant or misleading in the particular case (see Cronbach, 1976). The approach generally requires expertise of two kinds: Substantive, e.g., in educational program development and applications; and mathematical-statistical, e.g., in establishing presence and character of the stereotype and in accommodating the problem it generates. Similar approaches have been developed around particular evaluation designs, e.g., by Linn and Slinde (1977) in analyzing before-after evaluation designs, by Granger and Newbold (1974) in analyzing time-series data, by Linn (1976) in analysis of the use of standardized test norms, by Glaser (1973) in evaluating criminal offender programs, and by others.

Multiple Analytic Models and Methods of Analysis

Regardless of which broad approach one uses--checklists, stereotypes, comparative assessment--the problem of obtaining numerical estimates of the effect of a program remains. That is, the literally stated threats to validity, our understanding of how people behave in the absence of a project or program, must be translated into mathematical form to uncover the size of the program effect and to attach some level of confidence to the estimate. Having capitalized on substantive experts to learn about qualitative reality, the analyst must then match the mathematical models, which underpin any data anlaysis, with that reality.

The analyst faced with the prospect of estimating program effects based on nonrandomized data is, with some exceptions, not in an enviable position. There is a wide choice of methods which might be exploited, to be sure. For each, however, there is a set of assumptions which must be met in order for the analysis to be valid--i.e., to produce an unbiased estimate of effect. It is, in many cases, difficult to check assumptions, and for naively designed field surveys checking may be impossible.

The assumptions common to most methods bear on the properties of measurement and of the statistical models which underpin analysis. Most often, a piecewise approach to verifying their tenability is warranted. Scales of measurement in many analyses are supposed to have interval properties, for example. Roughly speaking, this means that the validity of measurement at any given point in the scale is as good as validity at any other point.

Attrition from treatment and comparison groups, if it occurs, is usually supposed to be random or at least predictable on the basis of simple statistical models. The fact that a program is in place, and has additive effects rather than (say) multiplicative effects, also must be verifiable both at the structural level and at the level of the individual participant. The structure of variability for any given analysis must usually have a well-specified form and evidence on the matter can determine which analysis is appropriate.

In some cases, there may not be a method which can be used without violating some assumption. There are two strategies which can then be used. One is to use the results of a sensitivity analysis, which is a test of how much a violation of a particular assumption will affect the outcome of the analysis. Occasionally this can be done through algebraic derivations, but usually it is done using simulation of data in which the assumption is violated. For example, most analysts are aware that violating the assumption of normal distribution of residuals has little effect on the statistical tests in analysis of variance. This was established through simulations of data in which various different distributions of data were used.

The second, related strategy is to try various methods of analysis, each of which involves a different set of assumptions. Under some circumstances, the pattern of results will enable the researcher to deduce not only the probable program effect, but also which assumptions were actually violated. Utilization of this method requires that the evaluator be able to list every assumption of each analytic method.

5. Recent Developments on Uniform Structures for Reporting on Evaluations

The guidelines for assaying quality of evaluations developed by the U.S. Agency for International Development, the U.S. General Accounting Office, UNESCO, and others constitute a vehicle for structuring evaluation reports. Uniform reporting has been helpful in summarizing project evaluations, in synthesizing results for program evaluation in sector level and country level analyses.

That there is some need for more orderly reporting is clear. Discussions at the 1979 UNESCO conferences on use of evaluation techniques in Asia stress the point (Ramalinga Iyer and others, 1979). Our own reviews of reports generated in AID evaluations and of the material available from AID's computerized archive system reiterate the need for rather more sophisticated and orderly reporting of evaluations.

The structure for reporting adopted here is an extension of one developed earlier in a state of the art monograph on social experimentation (Riecken, et al., 1974). It is related to more recent varieties of guidelines for reporting being developed by the U.S. General Accounting Office, and other groups (see the bibliography, on standards and guidelines).

In the following remarks, the structure of the cases is described first, followed by a more elaborate checklist for determining adequacy.

Structure for Reporting

The topics covered in each case study of an evaluation can be described in terms of the questions implied by each topic. The first three items catalogued below are preliminary; the next six items cover fundamental components of an outcome evaluation.

Background. What is the general nature of the social problem which is supposed to be addressed by the program under examination? In what sense is an outcome evaluation justified? Which organizations or individuals are principals in the evaluation?

General Objectives. What are the evaluation goals and how are they related to program objectives?

Program Goals and Target Groups. What is the specific rationale for the program? That is, how is the program expected to ameliorate the problem under attack? What group of individuals, or institutions, is the main target for the program? What plans, operations, staff, or material are necessary for the program's implementation?

Site and Communities of Interest. How was the site for the program or the pilot test selected? Which groups, organizations, or individuals exercise strong interest in or influence on the evaluation?

Once said, it is obvious that the choice of site for a program and its evaluation and the communities with an interest in the evaluation can exercise a strong influence on the quality of evaluation. Often, one cannot choose the site so as to reduce managerial, political, or institutional problems and that is the situation for most case studies considered here. In other cases, however, there is some choice. For example, the evaluators involved in the Nicaraguan radio mathematics experiment examined several countries before introducing the program to establish where such an experiment would be most appropriate and feasible. Most of the cases reported here also involve individuals or groups with a strong interest in understanding a new program's effects and that interest and support accounts partly for their quality. Other evaluations characterized by strong opposition or indifference have of course failed for those reasons. Some British tests on alternative methods for rehabilitating delinquents (Clark & Cornish, 1972) and at least one Guatemalan test on radio instruction (Hornik et al., 1973) were abandoned, for example, in the face of insurmountable obstacles to their continuation.

Evaluation Objectives, Design, Execution. What are the specific objectives of the evaluation? What design has been adapted to permit reasonably confident estimates of program effect to be made? What checks on proper execution of the design have been established?

The idea of conscientious design of evaluation, and the need to employ a design which is most likely to produce fair and interpretable estimates of program effect is fundamental. The evaluation plan should be explicit and documented before the fact. The failure of early evaluations of some programs in the United States, for example, has led to the U.S. General Accounting Office's dedicating notable attention to evaluation design in its reviews of the quality of evaluations and in construction of general guidelines on assessing evaluations. Just as programs may fail to be implemented adequately, evaluations may fail to be implemented in accord with evaluation design. Establishing that the design is executed then must be part of the evaluator's responsibility. If randomized assignment in experiments is planned, then the process should be monitored to assure that it is carried out. Two of the case studies presented later, the Canadian evaluation of a nurse-practitioner program and the Nicaraguan experiments on radio instruction, for example, take care to present evidence that the randomization procedure was carried out according to plan.

Measurement of Program Implementation. Recognizing that no new program ever conforms perfectly to plan, what mechanisms have been developed to assay level of implementation?

There's not much point to completing an outcome evaluation unless the program is actually in place, of course. The measures of program implementation may be very simple. A site visit, for instance, may be sufficient to understand that appropriate staff have been hired, that their activities are well documented and can be independently verified, and that individuals who are alleged to have received service actually did so. In the more intensive outcome evaluations, the measurement scheme is often more elaborate. The Canadian assessment of the effectiveness of nurse-practitioner programs, for example, included time and motion studies of activities in a conventional medical practice, daily journals on contacts between patients and medical staff, and ratings by independent observers on the observable adequacy of services provided.

Measurement of Outcomes and Other Characteristics of the Target Groups. Which specific characteristics of participants in alternative programs and of members of comparison groups are likely to be affected? How will these characteristics be measured? What mechanisms for assessing the quality of measures have been set up? How often were such measures undertaken? Are measures of preprogram characteristics appropriate, feasible, reliable?

Most programs have a variety of effects on their participants and, partly as a consequence, many field tests involve measurement of a variety of outcome variables. The better evaluations are usually characterized by pre-experiment development of measures and side studies of their reliability and validity. See, for example, the case studies on an educational enrichment program in Colombia and the radio mathematics program in Nicaragua. The evaluation may also involve attaching priorities to the outcome variables so as to make analysis more orderly, and to assure that the importance of various outcome variables for different groups with an interest in results is clear. The acquisition of other information or program participants before their entry to the program and on comparison groups is, of course, often necessary to assure comparability of groups, to increase the experiment's power to detect subtle program effects, and for other reasons. Most of the cases given here illustrate the point. The development and use of measures of outcome which are sensitive to the program's influence, and the design of side studies and monitoring systems to assure validity are discussed more fully in Riecken et al (1974).

Results. What kinds of data analyses were undertaken to estimate relative program effects? How is quality of the resulting estimate assessed? What inferences were drawn?

Most outcome evaluations involve a comparison between a group receiving services from a new program and those who receive services from an alternative or receive no services at all. The logic is normally supported with formal statistical tests which help one to understand whether the difference between the two groups is stable or merely a chance difference which could have occurred in the absence of any new program at all. The more conscientious evaluations also take care to examine reasons why a difference appears (or fails to appear) apart from the new program's influence. That is, the examination involves enumerating competing influences on the outcomes and an appraisal of the extent to which the outcome differences may be attributable to peculiar preprogram differences between groups or to some other influences. The idea is to assure that a declaration that a program effect is large is credible. As we suggested earlier, the process of establishing that an outcome is indeed attributable to a program's influence can be simplified through evaluation design and, in many cases, high quality design and execution is the only way to assure clear results.

Archive and Secondary Aanlysis. Have provisions been made to store the statistical data and to make it available for reanalysis by professionals not associated with the original program developers and evaluators?

In large scale evaluations, it is often reasonable to store resulting statistical data and to make it available for secondary analysis. Routine reanalysis is justified at times because, regardless of the integrity and skill of the original evaluator, independent analysis can lend credibility to the findings, help to identify errors in analysis, and so on. To the extent that the data are expensive, despite their benefits in understanding program effects, making the data available to the community of researchers for reanalysis is economical. A good deal of such data can be used for pedagogical purposes, for testing new hypotheses, and the like. Secondary analyses have been undertaken to assay credibility of the original evaluations of children's television programs, e.g., by Cook et al (1975) and by Cook and Conner (1976) on Sesame Street in the United States, Mexico, and elsewhere. They have been employed to understand the risks of trying to estimate program effects from simple one-time surveys, e.g., the Mosteller-Moynihan reanalyses of the Equality of Educational Opportunities Survey in the United States. It is efforts such as these which help to reinforce still earlier efforts to create cooperative social science data banks for storage of statistical data, i.e., records which have been stripped of the respondent's identification. Operational archives of this sort have been developed in Denmark (Danish Data Archive, 1976), Holland (Royal Netherlands Academy of Arts and Sciences, 1974), and Belgium (Université Catholique de Louvain, 1976). In the United States, the National Archives and several independent university-based facilities fulfill a similar mission.

Managerial Characteristics. Managerial features of the evaluation need to be reported, perhaps in appendices for a variety of reasons. Understanding how to allocate evaluation resources or how to budget for evaluations is difficult in the absence of information on precedent. And routine reporting on the topic can serve to be helpful in this respect. Such information includes details about why decisions about key aspects of evaluation were made, sources and nature of funding, description of the process of evaluation design and time schedule for the evaluation, manpower needs, budget, and costs of the evaluation.

Full Topical Checklist for Assessing Completeness and Quality of Outcome Evaluations

This checklist has been developed by Boruch, Rindskopf, and Rezmovic to facilitate the process of deciding whether an evaluation to estimate the effect of a social program is complete and of high quality. There are two columns on the checklist: Completeness/Quality and Responsibility. Completeness of an evaluation is assessed merely by noting whether an item in the list is, in fact, covered in the reports or other documentation. If the item is not relevant for a particular evaluation, the reader would merely put "NR" in the space. In case a more complete quality assessment is desired, some sort of numerical scale could be used, with "0" meaning the item was not covered, and the numbers 1 through 5 (or any other number sequence) used to indicate variations in quality from low to high. The column labelled "Responsibility" is included to assure that if an item is relevant, a particular individual or identified group was assigned to carry out the action. This is important for checking out why there might have been breakdowns in execution of the plan.

To the extent that a given evaluation actually considers relevant items in the checklist, it is good. But it is important to recognize that establishing the degree to which each item in the list is accommodated, determining how well plans are laid, is not simple. No checklist, including this one, will be adequate to the task. At its worst, the oversimplification implied by a checklist will be misleading.

Many items are not a matter of common practice in field evaluations of social programs. But it is our judgment, based on high quality evaluations by others and secondary analysis of them, that all items are a matter of good practice, are legitimate, and are usually feasible. There will be exceptions, to be sure, and these exceptions are among the main reasons for the Not Relevant column and for the examples.

The proper use of this checklist will depend heavily on the qualifications of the user: sometimes only wide experience in social research can enable a critic to make sound judgments about the severity of certain faults in design and execution. Criticism can be made at several different levels:

- Completeness and perfection: Was every item covered and covered well?
- Were plausible, not just possible, defects detected and corrected?
- Is the bias due to failure to detect and correct for defects likely to be serious?

Notice that increasing degrees of sophistication are required to make judgments in each successive case. The less sophisticated critic, who can only make judgments of the first kind, may reject findings which are perfectly reasonable because he was unable to judge the severity of the faults. In a brief checklist and manual such as this, we cannot hope to impart the knowledge necessary to make such judgments. They come only with extended training and experience.

Topical Checklist

for Outcome Evaluations

	Completeness Quality
1. Background	
. Nature of the social problem	_____
. General character of the program	_____
. General character of the evaluation	_____
. Sponsors and principals	_____
2. General Objectives	
. Program objectives	_____
. Evaluation objectives	_____
3. Program	
. Program goals and rationale (Evaluability)	_____
. Selection of program participants and target group	_____
. Planned program operations	_____
. Major elements of the program	_____
4. Site and Communities of Interest	
. Rationale for choice of test sites	_____
. Communities of interest and their cooperation	_____
5. Evaluation Objectives, Design, and Execution	
. Specific objectives of the evaluation	_____
. Evaluation design, including the logical, statistical model underlying design	_____
. Statistical power of design	_____
. Sample design and sample selection	_____
. Checks on execution of evaluation design and sampling	_____
. Threats to inferential validity characteristic of the evaluation design chosen	_____
6. Measures of Program Implementation	
Indicators of program presence	_____
. Indicators of program implementation level	_____
. Quality control on measurement	_____

7. Measurement of Outcomes and Other Characteristics

- Pre-program measurement: choice, justification, and quality control _____
- Post-program measures of outcome _____

  - .. rationale for choice _____
  - .. methods of measurement _____
  - .. quality control: reliability, validity, response rate _____
  - .. incentives for response _____

8. Data Analysis and Results

- Data analysis plan and rationale for it _____
- Formal statistical tests and models _____
- Competing analyses and threats to validity of conclusions _____
- Assumptions _____
- Statement of results _____

9. Archive and Secondary Analysis _____

- Plans for data storage _____
- Policy on and plan for independent reanalysis of data _____

10. Management Information

- Rationale/Justification for evaluation _____
- Sources and nature of funding _____
- Major decisions in evaluation design _____
- Time schedule for evaluation, projected and actual _____
- Manpower requirements _____
- Budget and expenditures for evaluation _____

4. On Human Factors in Program Evaluation

Evaluation often engenders anxiety among those whose program is being evaluated. It is equally obvious that maintaining honesty or integrity in evaluation can be difficult whether one is talking about evaluation, or demagoguery, or fraud. If it were not difficult to exhibit, integrity would not be regarded as a virtue. Permit me to presume here that the objective in principle is to do a job with integrity despite vague fears, political pressures, and similar influences.

I prefer not to espouse the virtue piously, but would like to embellish on a theme implied by an old evaluator of sorts, Mark Twain. He defined an ethical man, one with the integrity of the sort meant here, as a Christian holding four aces. We cannot always obtain the aces, but we can often negotiate for better cards, i.e., options which help to assure that if we do a job well, that we will not suffer needlessly the consequences of doing so.

Option 1. Posture at the Policy, Management, and Oversight Levels of Government.

There is some argument for the view that administrators of new and innovative projects ought not be judged on the basis of the success of the program for which they are responsible. For, as David Levintow, Donald Brown, and others have suggested at the recent AID conferences on evaluation, most AID projects are high risk ventures. And their failure is very often if not always beyond control of an individual or government. It is important to understand why we fail. Program managers and their staffs, then, should be judged on the quality of evidence bearing on a program, regardless, of whether one finds the program itself is a success. To be effective, that view would have to prevail at Congressional and administrative levels, and many readers will recognize that it also runs against social norms. That means if this posture is adopted generally, it will take a long time to become routine, perhaps as long as it took for the U.S. Census Bureau to infuse integrity routinely into U.S. censuses - 100 years.

For particular classes of innovation, it may be possible to employ the strategy immediately. And in others, some posture which provides rewards to both quality of evidence and quality of programs is warranted.

Option 2. Design of New Evaluations

It is sometimes sensible to accommodate fear of evaluation through design of evaluations. One of several simple ways of doing so is not to evaluate program A by comparing it to nothing: Rather one ought to compare A to B. "Nothing" may not be a politically viable option if A fails. Indeed, it's prudent to compare against B if one believes in planning: if A and B are equally effective, then one has a contingency plan, choosing B if it's less expensive, and recognizing the high risk of any innovative social program.

The difficulty with the option of course is that we may lack the imatination or resources to invent and execute B.

Option 3: External Review

One way to assure that incompetent evaluations and competent evaluations are properly labelled as such is to employ external reviews of evaluations. The tactic is consistent with the aims of the U.S. General Accounting Office and other agencies with an interest in quality and standards of evidence. It is consistent with the recent trend toward secondary analysis of program evaluation data, conducted by independent academic institutions. The latter option has been used, by among others, the U.S. Office of Education, the National Institute of Education, the Law Enforcement Assistance Administration, and other agencies in the United States. A variant on the tactic has been tried by Safraz Qureshi of Pakistan in reviewing evaluations sponsored by his government.

This option cannot assure directly that evaluations done with integrity will be rewarded. It should make it more likely, however, that the poor evaluations are not rewarded.

Option 4: Joint Dissemination and Review Panel Approaches

Consider a review board with clearly defined standards for examining the quality of evaluations, and which examine quality in response to a request from the program manager. A main objective of this panel or board is to officially verify that evidence is good and the program, if effective, deserves to be disseminated. Further, such a seal of approval can become a device for leveraging more money for similar projects from an agency. Indeed, panel approval may be a precondition for obtaining new money for diversifying, expanding, and so on.

In fact, such a system has been operating with some success by two domestic agencies in the U.S. The joint panel reviews educational products, basing review solely on evidence which conforms to well articulated standards. The approval makes them eligible for specially budgeted money earmarked for expansion, dissemination, and other purposes. The eligible programs compete for additional funds but with less competition than normal and more likelihood of success. About 30% of projects going before such a panel in the U.S. pass muster; this is up from 10% from the experience of 6-8 years ago.

Option 5: Evaluation as Research or as Management Tool

Labelling evaluation as a "research tool," the results of which are used for no fiscal decisions, rather than a management tool may ameliorate some problems. This option runs directly against the current thinking of GAO, OMB, and other U.S. agencies. Nonetheless, it is a posture which can indeed be taken in some studies. So, for instance, the Cali, Colombia, experiments on the impact of education and nutrition were undertaken by a research operation, the Human Ecology Research Institute, whose staff included two American university researchers, two Colombian researchers, and Colombian staff. It was undertaken with public and private foundation funds, was designed to conform to research standards of high quality evidence. The U.S. AID supported Nicaraguan experiments in radio-based mathematics instruction is of the same sort. The

remarkable feature of each is their treatment in principle and practice as research, potentially useable, but research nonetheless, free from untoward political pressures and no clear prior linkage to management decisions and the resistance to evaluation which decisions might engender.

Option 6: Reorganizing Mislabelling and Deception.

Bureaucratic recognition of mislabelling and deception is generally a high art form. It is not yet well developed in the evaluation arena, and it should be. To take the simplest case, descriptive surveys, needs assessment surveys, studies of management and operations, and the like are labelled "impact evaluations" or implied to be sufficient for impact estimation. The former are easy by comparison to impact studies, they differ in function and use, and they must not be confused. To the extent that they are, people who do a fine job on difficult enterprises may be inadvertently hurt by people doing a good job on easier tasks.

Option 7: Monitoring and Evaluating Use of Evaluation

One of the major concerns registered by program managers in AID at 1980 African meetings on evaluation, by program managers in local and regional education agencies in earlier meetings on U.S. Office of Education sponsored programs, and elsewhere concerns the uses of evaluation. Evaluations are supposed to be used after all, if they are good. It is in the program manager's interest to understand probable use, and moreover to protect oneself from ingenuous use.

Ingenuous use is possible of course. Many managers, legislative staffers, and so on have little understanding of the quality of evidence and so may naively rely on poor data. That result may affect those with good evidence negatively. Perhaps more important, the quality of utilization will vary depending on the experience and wit of the user. It is simply not easy to capitalize on evidence easily in the interest of making decisions or setting policy. One may, for example, decide on the basis of early evaluation results that a project is ineffective when longterm results can show that the program is effective. One may decide that the project in site 1 is ineffective when the project could or would work at site 2. And so on.

Some mechanisms must be invented to encourage, assure, and monitor, the high quality use of evaluation results. Several suggest themselves.

(a) Training sessions, workshops, etc. might be useful; this may include case study approaches as in Northwestern's or Harvard's MBA programs.

(b) Invention of cheap monitoring devices for identifying instances of information use, e.g. periodic telephone calls or cables to agency staff to ask whether information has been used, how it was used and who used it, and verification of the use.

(c) Development of case studies to illustrate ingenuous use and appropriate use.

There is some need for applied research on: "appropriate" and "inappropriate" use; delayed vs immediate uses; quality and quantity of use; types of use; and so on.

6. Common Questions and Partial Answers

This section describes some of the questions tendered by participants in AID's African conferences on evaluation of development programs. The responses here provide partial answers

1. Question: The effects of some development programs will only be detectable in 8-10 years. How can one evaluate under these conditions?

Consider first that there are large classes of programs which do have short term effects. Some effects of education programs for instance ought to be detectable immediately after a course. Increased arithmetic ability should be detectable following the course, though longterm effects of arithmetic on money handling in hotels, consumer buying habits, etc. will not be detectable for some time. Similarly, nutritional supplements will have some effects which detectable early, and others which are only detectable later.

The point is that one should not wait a long period to measure some effects. For long-term effects of some programs are not possible without some short-term effects.

Second, the statement that some programs have only long-term effects is a claim which may or may not be true. In point of fact, theory and evidence are often insufficient for making confident guesses about when program effects will appear. There is simply no coherent data on the matter for innovative programs (otherwise they probably wouldn't be regarded as innovative). The absence of data makes the claim less persuasive than it might otherwise be.

Third, if we are willing to admit that we are fairly ignorant about when effects will appear, then at least two tactics are sensible. One can, for example, measure outcomes periodically after the program's implementation to assure that regardless of when effects appear, the gradual nature of appearance is well understood. The measurement is made under the proviso that judgments about "success" are not legitimate for some time. One may also choose to measure direct and temporary as well as long-term effects to assure that immediate behavioral changes which are necessary for long-term changes do indeed appear.

2. Question: Are there situations in which one is reasonable not to try to estimate program effects?

Yes, there are lots.

One obvious class is charitable enterprise and related relief projects. There is a presumption that the work is good and regardless of whether the presumption is wrong, it is often sufficient to justify programmatic efforts on that presumption.

A second obvious class of projects for which impact evaluations may at times be unnecessary are those in which mere delivery is sufficient for justifying the resources. Examples here include the mundane: We believe teaching charity to children has long-term effects, and regardless of its real positive and negative effects, we may justify the activity solely in terms

of that belief, that article of faith. Similarly, one might choose to be satisfied with delivery of nutritional supplements to mothers as a basis for judging project "success" regardless of actual impact on their children. One may make the choice on the grounds that supplements are known to be good and their delivery is a sufficient good to justify resource expenditure. Similar presumptions justify investments in college education, music and literature, and so on. The presumption may be shortsighted: Supplements once delivered may not be consumed or they may be consumed in inappropriate ways, (e.g., mixing supplements with contaminated water).

The point is that the political, intellective, or religious presumption that delivery of A is a good, sometimes removes the need for impact evaluation. If one is unwilling to make the presumption, then impact evaluation may be warranted.

A third class of situations in which impact evaluations may not be warranted are those in which no one is interested in the results.

3. Question: Agencies such as AID have very little control over their projects in the sense that host countries exercise a great deal of influence. How can one evaluate well without more control?

The answer is that there are options, but no solution is equally appropriate for all settings. The options include:

(a) Host country involvement in evaluation as well as in project implementation, notably host country research and census units which are capable of cooperation and are competent.

(b) Selection of projects for elaborate evaluation solely on the basis of feasibility of evaluation, i.e., evaluate only those over which one has great control. The remaining projects receive less expensive, therefore more speculative, evaluation.

(c) Evaluate only those components of a complex program or project over which one has control, rather than evaluating the entire complex program.

(d) Stage the evaluation so as to obtain as much data as early as possible, and in such a way as to permit clear disengagement at several points in the process.

4. Question: What is the correct sample size to use in surveys?

Apart from recommending that one hire a reputable statistician to do the work, there is no simple answer to this question. The classical theory for choice of sample size hinges partly on the precision one would like in estimating characteristics of the target population, partly on cost. More recent theory for choice incorporates information about reliability of responses made to inquiry and probable nonresponse, i.e., nonsampling error.

The more fundamental and older question for managers is whether one ought to take an opportunistic sample, i.e., a sample of convenience, or a formal systematic sample. To illustrate the problem, consider the eleventh century rabbinical argument over this choice, where the need to choose originated from the need to estimate how much one should tithe. One might just grab a handful of olives from the nearest of a group of baskets to judge the general quality of the crop and tithe on the basis of this opportunistic sample. Or, one may formally mix the olives well, and develop a more formal, systematic, and time consuming sample. The rabbis' resolution of the issue is not much different from what contemporary analysts might achieve: if the demand for the estimate is rabbinical in origin, then one may take a sample of convenience. If it is biblical in origin, then one must be much more careful.

The more general point is that samples ought to be carefully designed if the use to which results will be put is important, and it need not be carefully designed otherwise.

If the manager decides that the problem is important enough to justify a formal survey, then the task of sample design ought to be given over to a competent statistician. To illustrate the statistician's thinking in the simplest case, consider the following example.

Suppose one's interest lies in estimating the proportion of tuberculosis victims in a city. The statistician will first make a guess, based on expert advice as to what proportion of the target population is so affected, say 50%. Either theory or readily available tables can then be used to establish the size of sample necessary to be very confident that the actual population proportion lies in a confidence interval. The intervals for various sample sizes and a guessed value of 50% to rate in the population are:

Sample Size (N)	Confidence Interval
25	.30-.70
100	.40-.60
400	.45-.55
900	.47-.53

The intervals narrow as sample size increases, implying a more precise estimate of the population parameter. But there is a point of diminishing returns, e.g. boosting sample size from 400 to 900 does not affect the confidence interval remarkably.

This is a rudimentary case of simple random sampling. The statistician may wish to improve precision of the estimate by designing a stratified sample. He or she may recognize that particular types of people with TB are of interest, implying that cross-tabulations are warranted, and that larger sample sizes are needed to obtain precision in detailed cross tabulations. If one attempts to accommodate nonsampling error (e.g., random unreliability in response), further increases will be necessary.

5. Question: Is there a special professional field of evaluation?

The field is new but growing rapidly, judging by the creation of new professional organizations and the issuance of journals.

Consider, for example, the professional organizations:

- Evaluation Research Society with 600 members
- Council for Applied Social Research with 300 members, most of whom do evaluative work
- Evaluation Network (with over 200 members, specializing in educational evaluation)

and pertinent divisions of the American Educational Research Association, the American Psychological Association, and others. ERS and CASR have international memberships, and are less than five years old. Existing organizations whose professional meetings place in developing countries (e.g., India, the Phillippines) such as the International Statistical Institute, run professional meetings at which evaluative research is reported regularly.

New journals dedicated to the field include:

- Evaluation Quarterly (Sage, Beverly Hills)
- Educational Policy and Evaluation (American Educational Research Association)
- Evaluation (Minneapolis)
- Evaluation Studies Review Annual (Sage)

and others. A compilation of such journals which provides addresses, editorial and subscription policies, is available in a report by Boruch (1979).

New university programs for evaluation have been cataloged by Ross Conner of the University of California at Irvine and John Morrell of Hahnemann Medical College, for the Evaluation Research Society. The catalog is available from the authors.

6. Question: How does one locate sources of technical assistance for evaluations in developing countries?

Finding competent assistance is difficult in evaluation as in any new arena of serious intellectual inquiry. There are several options.

It's sensible to read the evaluation literature for instance. An author who manages to say something intelligent in print is more likely than one who hasn't done so to provide competent services. If the problem at issue is controversial, one might hire both the author and the individual he or she suggests is the most competent opponent or critic of the author's views.

A major local source for technical assistance in the census bureau of the country in which the evaluation takes place. Many of the professional staff members are trained in the U.S. or Europe. The network of survey researchers in such bureaus is small enough to obtain informed recommendations about specific individuals or agencies. Most professional staff are members of professional organizations such as the International Statistical Institute or the American Statistical Association. The papers they present at annual meetings of the organization are typically published in proceedings and the ppaers serve as one vehicle for judging quality of work.

Other major sources include social and economic research units, human ecology research institutes, and other institutional research groups. They are not uncommon in India, Colombia, Malaysia, Pakistan, and elsewhere, though quality of work varies considerably. The Human Ecology Research Institute in Cali, Colombia has mounted some of the finest experimental tests of education programs available to date. I know of no formal list of such organizations.

Universities also consitute a resource for technical assistance. The resource may be limited at times: some of the University of Nairobi faculty are not in good favor with the government of Kenya, for example. But this need not prevent their participation in independently supported projects; indeed obtaining the assistance of high calibre people in and out of favor may enhance quality of the evaluation. We are aware of no formal examination of this topic however.

7. Question: Why are we learning all this theory?

The criticism implied in the question is that "all this is theoretical, but not the real world."

In either case, the choice of the word theory is wrong if (a) all illustrations are real and (b) the lecturer is actually not a theoretician.

The more appropriate reference is to the distinction between principle and practice. That is, some evaluation designs are as a matter of principle good in yielding good evidence. But they are not always possible.

(INTERIM REPORT)

On Impact Evaluations in Developing Countries. II

A Bibliography on Methods Pertaining to Impact Evaluation, with Special Attention to Developing Countries and to Recent Methodological Developments

Robert F. Boruch

Northwestern University

Report No. A-92/Appendix

Pursuant to: Contract with the Agency for International Development. Earlier research on related methodological topics was supported by National Science Foundation and the National Institute of Education.

Date: January, 1980

Methodology and Evaluation Research
Psychology Department
Northwestern University
Evanston, Illinois 60201 USA

48

APPENDIX

PAGE

1. Logic and Fundamental Elements of Evaluation
   - 1.1 Standards, Guidelines, and Glossaries in Evaluation
   - 1.2 Issues: Fundamental and Otherwise
2. Evaluability
3. Fundamental Aspect of Applied Social Research: Questionnaires, Samples, Evaluation Designs, Analysis
4. Qualitative Methods
5. Qualitative and Quantitative
6. Surveys and Survey Samples
7. Measurement, Measurement Error, and Nonsampling Error
   - 7.1 Measurement of Sensitive Traits: Randomized Response
   - 7.2 Dual System Estimates
8. Perceptions and Opinions in Evaluation
9. Measurement of Process and Implementation
10. Sensitivity of the Evaluation Design and Relation to Sensitivity of Measurement.
11. Change Scores and Before - After Designs: New Developments
12. Logic, Use and Abuse of Regression, Covariance, and Matching
13. Matching and Control Group Designs: New Developments
14. Regression and Analysis of Covariance with Errors of Measurement Misspecification of Variables, including Selection Problems
15. Bayesian Inference
16. Factor Analytic and Regression Approaches
17. Structural Equation Models
18. Randomized Experiments: Design
19. Regression - Discontinuity Designs
20. Time Series and Longitudinal Data Designs
21. Time Series Applications
22. Ex-Post Facto Designs
23. Count Data: Analysis
24. Pattern Analysis
25. Secondary Analysis
26. Meta Analysis

1. LOGIC AND FUNDAMENTAL ELEMENTS OF EVALUATION

Agency for International Development. Evaluation handbook. Washington, D.C.: USAID, 1976.

Agency for International Development. A field guide for evaluation of nutrition education. Washington, D.C.: Office of Nutrition, Technical Assistance Bureau, 1975.

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Campbell, D. T. Assessing the impact of planned social change. In G. M. Lyons (Ed.), Social research and public policies: The Dartmouth/OECD Conference. Hanover, N.H.: University Press of New England, 1975, pp. 3-45.

Campbell, D. T. Focal local indicators for social program evaluation. Social Indicators Research, 1976, 3, 237-256.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963. (Also published as Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1966.)

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Skokie, Ill.: Rand McNally, 1976.

Freeman, H. E. The present status of evaluation research. In M. Guttentag (Ed.), Evaluation Studies Review Annual, 1977, 2, 17-51.

Glaser, D. Routinizing evaluation: Getting feedback on effectiveness of crime and delinquency programs. Rockville, Maryland: National Institute of Mental Health, 1973.

Hoole, F. W. Evaluation research and development activities. Beverly Hills: Sage, 1978.

Mukherjee, B. Selected techniques of social programme evaluation: Some Indian experience. Council for Social Development, New Delhi: (53 Lodi Estate, 110003), December, 1979.

Riecken, H. W., Boruch, R. F., Campbell, D. T., Glennan, T. K., Pratt, J., Rees, A., & Williams, W. Social experimentation: A method for planning and evaluating social programs. New York, N.Y.: Academic Press, 1974.

Rossi, P. H. Boobytraps and pitfalls in the evaluation of social action programs. In F. G. Caro (Ed.), Readings in Evaluation Research. New York: Russell Sage, 1977.

Rossi, P. H., & Williams, W. Evaluating social programs: Theory, practice and politics. New York: Seminar Press, 1972.

Rossi, P. H., Freeman, H. E., & Wright, S. R. Evaluation: A systematic approach. London & Beverly Hills, Ca: Sage, 1979.

Rossi, P. H., Wright, J. D., & Wright, S. R. The theory and practice of applied social research. Evaluation Quarterly, 1978, 2(2), 171-192.

Rutman, Leonard. Evaluation research methods: A basic guide. Beverly Hills: Sage Publications, 1977.

Rutman, L. Formative research and program evaluability. In L. Rutman (Ed.), Evaluation research method: A basic guide. Beverly Hills, Ca.: Sage, 1977, pp. 57-72.

Sohm, E. D., & Bertrand, M. Initial guidelines for internal evaluation systems of United Nations Organizations. Geneva, Switzerland: UN Joint Suspection Unit, February, 1979.

Soumelis, C. G. Project evaluation methodologies and techniques. Paris: UNESCO, 1977.

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

Weiss, C. H. Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

1.1 Standards, Guidelines, and Glossaries in Evaluation

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

Davis, H. R., Windle, C., & Sharfstein, S. S. Developing guidelines for program evaluation capability in Community Mental Health Centers. Evaluation, 1977, 4, 24-34.

Gonzalez, M. E., Ogus, J. L., Shapiro, G., & Tepping, B. Standards for discussion and presentation of errors in survey and census data. Journal of the American Statistical Association, 1975, 70(351, Part II), 5-23.

Gordon, G., & Morse, E. V. Evaluation research. In A. Inkeles (Ed.), Annual Review of Sociology, 1975 (Vol. 1).

Hennigan, K., Flay, B., and Haag, R. Clarification of concepts and terms commonly used in evaluative research. In R. E. Klein, et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 387-432.

Joint Committee on Standards for Education Evaluation. Draft Standards for Evaluation of Educational Programs, Projects, and Materials. (Edited by Daniel L. Stufflebeam). Western Michigan University, Kalamazoo, Michigan, 1977.

McTavish, D. G., Cleary, J. D., Brent, E. E., Perman, L., & Knudsen, K. R. Assessing research methodology: The structure of professional assessments of methodology. Sociological Methods and Research, 1977, 6(1), 3-44.

Sohm, E. D., & Bertrand, M. Initial guidelines for internal evaluation systems of United Nations Organizations. Geneva, Switzerland: UN Joint Suspection Unit, February, 1979.

Tallmadge, G., & Horst, P. Designs for local educational program evaluations. RMC Corporation Report, 1977.

U.S. Agency for International Development. Evaluation handbook (Second Edition). Washington, D.C.: USAID, Office of Program Evaluation, 1976.

U. S. General Accounting Office. Evaluation and analysis to support decision-making. Washington, D.C.: U.S. GAO (PAD-26-9), December, 1975.

U. S. General Accounting Office. Assessing social program impact evaluations: A checklist approach. Washington, D.C.: U.S. GAO (PAD-79-2), October, 1978.

1.2 Issues: Fundamental and Otherwise

Cook, T. D., and McAnany, E. G. Recent United States experiences in evaluation research with implications for Latin America. In: R. E. Klein et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 39-76.

Hollister, R. Comments on "Recent United States experiences in evaluation research with implications for Latin America. In: R. E. Klein, et al (Eds.) Evaluating the impact of nutrition and health programs. In R. E. Klein et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 76-97.

Riecken, H. W. Practice and problems of evaluation: A conference synthesis. In R. E. Klein, et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 363-386.

Townsend, J. W., Farrell, W. T., and Klein, R. E. Special issues for the measurement of program impact in developing countries. In R. E. Klein, et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 99-126.

2. EVALUABILITY, FEASIBILITY, AND APPROPRIATENESS OF EVALUATIONS

Boruch, R. F., Anderson, P. S., Rindskopf, D. E., Amidjaya, I. R., & Jansson, D. Randomized experiments for evaluating and planning local programs: A summary on appropriateness and feasibility. Public Administration Review, 1979, 39(1), 36-40. Reprinted in H. E. Freeman (Ed.) Policy Studies Review Annual, 1978, 2, 226-236 (Beverly Hills: Sage).

Rutman, L. Formative research and program evaluability. In L. Rutman (Ed.), Evaluation research method: A basic guide. Beverly Hills, Ca.: Sage, 1977, 57-72.

Rutman, L. Planning useful evaluations: evaluability assessment. London, U.K. and Beverly Hills, Ca.: Sage, 1980.

Wholey, J. S. Evaluability assessment. In L. Rutman (Ed.), Evaluation research methods: A basic guide. Beverly Hills, Ca.: Sage, 1979, pp. 39-58.

Wholey, J. S., Scanlon, J. W., Nay, J. N., & Schmidt, R. E. If you don't care where you get to, then it doesn't matter which way you go. In G. M. Lyons (Ed.), Social research and public policies: The Dartmouth/OECD Conference. Hanover, N.H.: University Press of New England, 1975, pp. 175-197.

Wholey, J. S., Nay, J. N., Scanlon, J. W., & Schmidt, R. E. Evaluatic : When is it really needed? Evaluation, 1975, 2, 89-93.

3. FUNDAMENTALS OF RESEARCH: QUESTIONNAIRES, SAMPLES, EVALUATION DESIGNS, ANALYSIS

Babbie, E. R. Survey research methods. Belmont, Ca.: Wadsworth, 1973.

Groves, R. M., and Kahn, R. L. Surveys by telephone. New York and London: Academic, 1979.

Practical Concepts Inc. Manager's guide to data collection. Washington, D.C.: PCI (1730 Rhode Island Avenue NW), 1979.

4. QUALITATIVE METHODS

Britan, G. M. The place of anthropology in program evaluation. Anthropological Quarterly, 1978, 51(2), 119-128.

Britan, G. M. Experimental and contextual models of program evaluation. Evaluation and Program Planning, 1978, 1, 229-234.

Britan, G. M. Evaluating a federal experiment in bureaucratic reform. Human Organization, 1980, in press.

Campbell, D. T. Qualitative knowing in action research. Kurt Lewin Award Address Before the Society for the Psychological Study of Social Issues, Annual Meeting, New Orleans. To appear in Journal of Social Issues, 1975.

Knapp, M. S. Ethnographic contributions to evaluation research: The experimental schools program evaluation and some alternatives. In T. D. Cook and C. S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research. Beverly Hills, Ca.: Sage, 1979, pp. 118-139.

Mitchell, W. D. Finding out what happens in the field: The use of field logs. Evaluation Quarterly, 1979, 3, 427-445.

Wilson, S. Explorations on the usefulness of case study evaluations. Evaluation Quarterly, 1979, 3, 446-459.

Wilson, S. The use of ethnographic techniques in educational research. Review of Educational Research, 1977, 47(2), 245-266.

5. QUALITATIVE AND QUANTITATIVE

Barzun, J. Clio and the doctors: Psycho-history, quanto-history and history. Chicago: University of Chicago Press, 1974.

Boruch, R. F. Common contentions about randomized experiments for planning and evaluating social programs. In: R. F. Boruch & W. H. Riecken (Eds.), Experimental testing of public policy. Boulder, Colorado: Westview Press, 1975, pp. 108-145.

Boruch, R. F., & Gomez, H. Sensitivity, bias, and theory in impact evaluations. Professional Psychology, 1977, 8, 411-434.

Britan, G. M. The place of anthropology in program evaluation. Anthropological Quarterly, 1978, 51(2), 119-128.

Cohen, R. Warring epistemologies: Quality and quantity in African research. In W. M. O'Barr, et al., Survey research in Africa: Its applications and limits. Evanston: Northwestern University Press, 1973.

Cook, T. D., & Cook, A. F. L. Comprehensive evaluation and its dependence on both humanistic and empiricist perspectives. In R. S. French (Ed.), Humanists and policy studies: Relevance revisited. Washington, D.C.: George Washington University, Division of Experimental Programs, 1977.

Cook, T. D., & Reichardt, C. S. (Eds.). Quantitative and qualitative methods in evaluation. Beverly Hills, Ca.: Sage Publication, in press.

Reichardt, C. S., & Cook, T. D. Beyond qualitative versus quantitative methods. In T. D. Cook and C. S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research. Beverly Hills, Ca.: Sage, 1979, pp. 7-32.

Trend, M. G. On the reconciliation of qualitative and quantitative analyses: A case study. In T. D. Cook and C. S. Reichardt (Eds.), Qualitative and quantitative methods in evaluation research. Beverly Hills, Ca.: Sage, 1979, pp. 68-86.

14. REGRESSION AND COVARIANCE ANALYSIS WITH ERRORS OF MEASUREMENT IN THE INDEPENDENT VARIABLE, ERRORS IN VARIABLES, OR SELECTION PROBLEMS

Bentler, P. M., & Woodward, J. A. Non-experimental evaluation research: Contributions of causal modeling. In L. E. Datta and R. Perloff (Eds.), Improving evaluations. Beverly Hills, Ca.: Sage, 1979, pp. 71-102.

Berk, R. A., Hennessy, M., & McCleary, R. Descriptive distortions in covariance based statistics. Evanston, Illinois: Sociology Department, Northwestern University, 1977. (Available from R. A. Berk, Sociology, University of California at Santa Barbara.)

Boruch, R. F., & Gomez, H. Sensitivity, bias, and theory in impact evaluations. Professional Psychology, 1977, 8, 411-434.

Bryk, A. S., & Weisberg, H. I. The implications of nonrandom assignment in comparative studies involving growth systems. Proceedings of the American Statistical Association, Social Statistics Section. Washington, D.C.

Bryk, A. S., & Weisberg, H. I. Use of nonequivalent control group design when subjects are growing. Psychological Bulletin, 1977, in press.

Cain, G. G. Regression and selection models to improve nonexperimental comparisons. In C. A. Bennett and A. A. Lumsdaine (Eds.), Evaluation and experiment: Some critical issues in assessing social programs. New York: Academic Press, 1975.

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

Campbell, D. T., & Boruch, R. F. 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. Lumsdaine (Eds.), Evaluation and experiment. New York: Academic Press, 1975.

Cochran, W. D. Errors of measurement in statistics. Technometrics, 1968, 10, 637-666.

Cochran, W. G. Some effects of errors of measurement on multiple correlation. Journal of the American Statistical Association, 1970, 65, 22-34.

Cochran, W. G. Some effects of errors in measurement on linear regression. In L. M. LeCam, J. Neyman, and E. L. Scott (Eds.), Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics. Los Angeles: University of California, 1972.

13. MATCHING AND CONTROL GROUPS

Cochran, W. G., & Rubin, D. G. Controlling bias in observational studies: A review. Sankhya, 1973, 35(4), Series A, 417-556.

Hardin, E. On the choice of control groups. In M. E. Borus (Ed.), Evaluating the impact of manpower programs. Lexington, Mass.: Lexington Books, 1972.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68(5), 304-305.

McKinlay, S. M. The design and analysis of the observational study--A review. Journal of the American Statistical Association, 1975, 70, 503-520.

Miller, D. A. On the choice of control groups: Comment. In M. E. Borus (Ed.), Evaluating the impact of manpower programs. Lexington, Mass.: Lexington Books, 1972. Pp. 59-62.

Rubin, D. G. The use of matched sampling and regression adjustment to remove bias in observational studies (Technical Report No. 34). Office of Naval Research, Department of Statistics, June 24, 1971.

Rubin, D. B. Multivariate matching methods that are equal percent bias reducing, II. Maximums of bias reduction for fixed sample sizes. Biometrics, 1976, 32, 121-132.

Rubin, D. B. Multivariate matching methods that are equal percent bias reducing, I: Some examples. Biometrics, 1976, 32, 109-120.

Saxe, L., & Fine, M. Expanding our view of control groups in evaluations. In L. E. Datta & R. Perloff (Eds.), Improving evaluations. Beverly Hills, Ca.: Sage, 1979, pp. 61-70.

Thorndike, E. L. The influence of chance imperfections of measures upon the relationship of initial score to gain or loss. Journal of Experimental Psychology, 1924, 7, 225-232.

Thorndike, R. L. Regression effects in the matched groups experiment. Psychometrika, 1942, 7, 85-102.

6. SURVEYS AND SURVEY SAMPLES

Brown, R. B. Research and the credibility of estimates. Boston: Harvard University Press, 1969.

Drake, H. M. Research method or culture-bound technique? Pitfalls of survey research in Africa. In W. M. O'Barr, et al., Survey research in Africa: Its applications and limits. Evanston: Northwestern University Press, 1973.

Hursh-Cesar, G., & Roy, P. Third World surveys: Survey research in developing nations. New Delhi: Macmillan Company of India, 1976.

Murphy, M. N. Sampling theory and methods. Calcutta, India: Statistical Publishing Society, 1977.

Namboodiri, N. K. Survey sampling and measurement. New York: Academic, 1978.

O'Barr, W. M., Spain, D. H., & Tessler, M. A. Survey research in Africa: Its applications and limits. Evanston: Northwestern University Press, 1973.

Sparks, R. F., Genn, H. G., Dodd, D. J. Surveying victims. Chichester, U.K.: Wiley, 1977.

United Nations. Recommendations for the preparation of sample survey reports. Statistical Paper Series, C, No. 1, Rev. 2. New York: UN, 1964.

7. MEASUREMENT, MEASUREMENT ERROR, AND NONSAMPLING ERROR

Brooks, C. A., & Bailar, B. A. An error profile: Employment as measured by the current population survey. U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, 1978.

Cronbach, L. J., et al. The dependability of behavioral measurements. New York: Wiley, 1972.

Gruijter, D. N. M., & van der Kamp (Ed.). Advances in psychological and educational measurement. London: Wiley, 1976.

Hansen, M. H., Hurwitz, W. N., Marks, E. S., & Mauldin, W. P. Response errors in surveys. Journal of the American Statistical Association, 1951, 46, 146-190.

Kiresuk, T. J., & Lund, S. H. Process and outcome measurement using goal attainment scaling. In G. V. Glass (Ed.), Evaluation Studies Review Annual, 1976, 1, 383-399.

Krotki, K. (Ed.). Developments in dual system estimation of population size and growth. Alberta, Canada: University of Alberta Press, 1978.

Mason, W. M., Hauser, R. M., Kerchoff, A. C., Poss, S. S., & Manton, K. Models of response error in student reports of parental socioeconomic characteristics. In W. H. Sewell, R. M. Hauser, and D. L. Featherman (Eds.), Schooling and achievement in American society. New York: Academic Press, 1973, pp. 443-519.

Sudman, S., & Bradburn, N. M. Improving interview method and questionnaire design. San Francisco: Jossey-Bass, 1979.

7.1 MEASUREMENT OF SENSITIVE TRAITS: RANDOMIZED RESPONSE

Abernathy, J. R., Greenberg, B. G., & Horvitz, D. G. Estimates of induced abortion in urban North Carolina. Demography, 1970, 7(1), 19-29.

Barth, J. T., & Sandler, H. M. Evaluation of the randomized response technique in a drinking survey. Journal of Studies on Alcohol, 1976, 37(5), 690-693.

Berman, J., McCombs, H., & Boruch, R. F. Notes on the contamination methods. Sociological Methods and Research, 1977, 6, 45-62.

Brown, G. H., & Harding, F. D. A comparison of methods of studying illicit drug usage. HUMRO Technical Report 73-9, Alexandria, Va.: Human Resources Research Organization, 1973.

Brown, G. Randomized inquiry vs. conventional questionnaire method in estimating drug usage rates through mail surveys. HUMRO Technical Report 75-14, Arlington, Virginia: Human Resources Research Organization, June 1975.

Eriksson, S. A. Randomized interviews for sensitive questions. (Ph.D. dissertation) University Institute Statistics, University of Gothenburg, (Vasagatan 7) Gothenburg, Sweden, 1973(a).

Fidler, D. S., & Kleinknecht, R. E. Randomized response versus direct questioning: Two data collection methods for sensitive information. Psychological Bulletin, 1977, 84(5), 1045-1049.

Folsom, R. E. A randomized response validation study: Comparison of direct and randomized resporting of DUI arrests. Final Report, 2550-807, Research Triangle Institute, Chapel Hill, North Carolina, August 1974.

Goodstadt, M. S., & Gruson, V. The randomized response technique: A test on drug use. Journal of the American Statistical Association, 1975, 70, 814-818.

I-cheng, C., Chow, L. P., & Rider, R. V. The randomized response technique as used in the Taiwan outcome of pregnancy study. Studies in Family Planning, 1972, 3(11), 265-269.

Krotki, K., & Fox, B. The randomized response technique, the interview, and the self-administered questionnaire: An empirical comparison of fertility reports. Proceedings of the Social Statistics Section: American Statistical Association. Washington, D.C.: ASA, 1974.

Liu, P. T., Chen, C. N., & Chow, L. P. A study of the feasibility of Hopkins randomized response models. Proceedings of the American Statistical Association: Social Statistics Section, Part II. Washington, D.C.: ASA, 1976, 561-566.

Locander, W., Sudman, S., & Bradburn, N. An investigation of interview method, threat, and response distortion. Journal of the American Statistical Association, 1976, 71(354), 269-275.

O'Brien, D. M., Cochran, R. S., Marquardt, R. S., & Makens, J. C. Randomized response vs direct question in a mail vs personal interview consumer opinion survey. College of Commerce and Industry Research Paper No. 85, Laramie, Wyoming: University of Wyoming, (July) 1975.

Reaser, J. M., Hartsock, S., & Hoehn, A. J. A test of the forced alternative random response questionnaire technique. HUMRO Technical Report 75-9, Arlington, Virginia: Human Resources Research Organization, June 1975.

Shimizu, I. M., & Bonham, G. S. Randomized response technique in a national survey. Journal of the American Statistical Association, 1978, 73(31), 35-39.

Zdep, S. M., & Rhodes, I. N. Making the randomized response technique work. Public Opinion Quarterly, 1977, 41, 531-537.

7.2 DUAL SYSTEM ESTIMATES

Ahmed, Nazir, & Krotki, Karol. Simultaneous estimation of population growth: The Pakistan experiment. Pakistan Development Review, 1963, 3, 37-65.

Beaujot, R. P. The PGE/ERAD/ECP system of data collection in Africa and a comparison of its results with those of analytic techniques. In K. Krotki (Ed.), Developments in dual system estimation of population size and growth. Alberta, Canada: University of Alberta Press, 1979, pp. 145-152.

Brass, W. I. The use of existing data. In: J. C. Caldwell and C. Okonjo (Eds.), The population of tropical Africa. New York: Columbia University Press, 1968, pp. 172-178.

Brass, W. I., & Coale, A. J. Methods of analysis and estimation. In: W. I. Brass, et al. (Eds.), The demography of tropical Africa. Princeton, N. J.: Princeton University Press, 1968, pp. 88-139.

de Lamaz, Francois. The collection of demographic data in Francophone Africa and Liberia using the PGE/ERAD/ECP system. In K. Krotki (Ed.), Developments in dual system estimation of population size and growth. Alberta, Canada: University of Alberta Press, 1978, pp. 92-99.

Housni, El Arbi, Notzon, F., & Fichet, M. D. PGE/ERAD/ECP matching experiences in Morocco. In K. Krotki (Ed.), Developments in dual system estimation of population size and growth. Alberta, Canada: University of Alberta Press, 1978, pp. 122-140.

Marks, S., & Rumford, J. The 1974 Post-enumeration survey of Liberia--A new approach. In K. Krotki (Ed.), Developments in dual system estimation of population size and growth. Alberta, Canada: University of Alberta Press, 1978, pp. 189-202.

Rachidi, Mohamed. Some practical problems suggested by the application of the PGE/ERAD/ECP system in Morocco. In K. Krotki (Ed.), Developments in dual system estimation of population size and growth. Alberta, Canada: University of Alberta Press, 1978, pp. 113-118.

Vaidyanathan, K. E. The Egyptian study to measure vita rates: Some estimates by dual collection. In K. Krotki (Ed.), Developments in dual system estimation of population size and growth. Alberta, Canada: University of Alberta Press, 1978, pp. 104-109.

8. PERCEPTIONS AND OPINIONS

Scheirer, M. A. Program participants' positive perceptions: Psychological conflict of interest in social program evaluation. In L. Sechrest, et al. (Ed.), Evaluation Studies Review Annual, 1979, 4, 407-424.

Stipak, B. Citizen satisfaction with urban services: Potential misuse as a performance indicator. In L. Sechrest, et al. (Ed.), Evaluation Studies Review Annual, 1979, 4, 440-447.

Sudman, S., & Bradburn, N. M. Improving interview method and questionnaire design. San Francisco: Jossey-Bass, 1979.

9. MEASUREMENT OF PROCESS OR IMPLEMENTATION

Ahmed, Manzoor, & Coombs, P. H. Education for rural development: Case studies for planners. New York: Praeger, 1975.

Larkey, P. D. Process models of governmental resource allocation and program evaluation. In L. Sechrest, et al. (Ed.), Evaluation Studies Review Annual, 1979, 4, 277-283.

Moos, R. H. Evaluating educational environments. San Francisco: Jossey-Bass, 1979.

Parlett, M., & Hamilton, D. Evaluation as illumination: A new approach to the study of innovatory programs. In G. V. Glass (Ed.), Evaluations Studies Review Annual, 1976, 1, 140-157.

Patton, M. Q. Evaluation of program implementation. In L. Sechrest, et al. (Ed.), Evaluation Studies Review Annual, 1979, 4, 318-348.

Rossi, P. H., & Williams, W. Evaluating social programs: Theory, practice and politics. New York: Seminar Press, 1972.

10. SENSITIVITY OF THE EVALUATION DESIGN AND RELATION TO SENSITIVITY OF MEASUREMENT

Boruch, R. F., & Gomez, H. Sensitivity, bias, and theory in impact evaluations. Professional Psychology, 1977, 8, 411-434. Reprinted in L. Datta and R. Perloff (Eds.) Improving evaluations. Beverly Hills: Sage, 1979, pp. 139-170.

Cleary, T. A., Linn, R. L., & Walster, G. W. The effect of reliability and validity on power of statistical tests. In E. F. Borgatta and G. W. Bohrnstedt (Eds.), Sociological Methodology. San Francisco: Jossey-Bass, 1970.

Habicht, J. P., and Butz, W. P. Measurement of health and nutrition effects of large-scale nutrition intervention projects. In R. E. Klein (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 133-182.

Nicholson, W.. & Wright, S. R. Participants' understanding of the treatment in policy experimentation. Evaluation Quarterly, 1977, 1(2), 195-244.

Proctor, C. H. Testing for contextual effect. Proceedings of the American Statistical Association: Social Statistics Section Part II. Washington, D.C.: ASA, 1976, pp. 695-699.

11. CHANGE SCORES: BEFORE-AFTER DESIGNS AND ANALYSIS

Bereiter, C. Some persisting problems in the measurement of change. In C. W. Harris, Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963.

Cronbach, L. J., & Furby, L. How should we measure change--Or should we? Psychological Bulletin, 1970, 74, 68-80.

Hummel-Rossi, B., & Weinberg, S. Practical guidelines in applying current theories to the measurement of change. Catalog of Selected Documents in Psychology, 1975, 5, 226.

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82(3), 345-363.

Linn, R. L., & Slinde, J. A. The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 1977, 47, 121-150.

Lord, F. M. The measurement of growth. Educational and Psychological Measurement, 1956, 16, 421-437.

Lord, F. M. Further problems in measurement of growth. Educational and Psychological Measurement, 1958, 18, 437-454.

Lord, F. M. Large-scale covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 1960, 55, 307-321.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1969, 72(5), 336-337.

Lord, F. M. Statistical adjustments when comparing pre-existing groups. Psychological Bulletin, 1969, 72(5), 336-337.

McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.

O'Conner, E. F. Extending classical test theory to the measurement of change. Review of Educational Research, 1972, 42(1), 73-97.

O'Conner, E. F. Response to Cronbach and Furby's "How should we measure 'change'--or should we?" Psychological Bulletin, 1972, 78(2), 159-160.

Stanley, J. C. General and special formulas for reliability of differences. Journal of Educational Measurement, 1967, 4, 249-252.

Thorndike, E. L. The influence of chance imperfections of measures upon the relationship of initial score to gain or loss. Journal of Experimental Psychology, 1924, 7, 225-232.

Tuchfield, B. S. Some approaches to measuring change. In L. E. Datta & R. Perloff (Eds.), Improving evaluations. Beverly Hills, Ca.: Sage, 1979, pp. 103-110.

Tucker, L. R., Damarin, F., & Messick, S. A base-free measure of change. Psychometrika, 1966, 31, 457-473.

Werts, C. E., & Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 23, 17-23.

12. LOGIC, USE, AND ABUSE OF REGRESSION, COVARIANCE, AND MATCHING

Berkson, J. Are there two regressions? Journal of the American Statistical Association, 1950, 45, 164-180.

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

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

Campbell, D. T., & Boruch, R. F. 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. Lumsdaine (Eds.), Evaluation and experiment. New York: Academic Press, 1975.

Granger, C. W. J., & Newbold, P. Spurious regressions in econometrics. Journal of Econometrics, 1974, 2, 111-120.

Hardin, E. On the choice of control groups. In M. E. Borus (Ed.), Evaluating the impact of manpower programs. Lexington, Mass.: Lexington Books, 1972.

Hilton, G. Causal inference analysis: A seductive process. Administrative Science Quarterly, 1972, March, 44-54.

James, K. I. Regression toward the mean in uncontrolled clinical studies. Biometrics, 1973, 29, 121-130.

Miller, D. A. On the choice of control groups: Comment. In M. E. Borus (Ed.), Evaluating the impact of manpower programs. Lexington, Mass.: Lexington Books, 1972. Pp. 59-62.

McKinlay, S. M. The design and analysis of the observational study--A review. Journal of the American Statistical Association, 1975, 70, 503-520.

McCleary, R., et al. How a regression artifact can make any delinquency program look effective. In L. Sechrest, et al. (Ed.), Evaluation Studies Review Annual, 1979, 4, 626-652.

McLean, J. E., Ware, W. B., & McClave, J. T. How analysis of covariance can yield misleading results in educational experiments: A Monte Carlo study. Proceedings of the American Statistical Association: Social Statistics Section, Washington, D.C.: ASA, 1975, pp. 548-553.

Overall, J. E., & Woodward, J. A. Common misconceptions concerning the analysis of covariance. Multivariate Behavioral Research, 1977, 12, 171-186.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 1974, 66(5), 688-701.

Rulon, P. J. Problems of regression. Harvard Educational Review, 1941, 11, 213-223.

Smith, H. G. Interpretation of adjusted treatment means and regressions in analysis of covariance. Biometrics, 1957, 13, 282-307.

Thorndike, R. L. Regression effects in the matched groups experiment. Psychometrika, 1942, 7, 85-102.

Werts, C. E., & Linn, R. L. Analyzing school effects: How to use the same data to support different hypotheses. American Education Research Journal, 1969, 6, 439-447.

Wold, H. Causal inference from observational data: A review of ends and means. Journal of the Royal Statistical Society, 1956, 69 (Part I), 28-[illegible]9.

Cronbach, L. J., et al. The dependability of behavioral measurements. New York: Wiley, 1972.

Deegan, J. The consequences of model misspecification in regression analysis. Multivariate Behavioral Research, 11, 237-248.

DeGracie, J. S. Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. Ph.D. dissertation. Iowa State University (Ames).

DeGracie, J. S., & Fuller, W. A. Estimation of slope and analysis of covariance when the concomitant variable is measured with error. Journal of the American Statistical Association, 1972, 67(340), 930-937.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6(3), 383-401.

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

Feldstein, M. Errors in variables: A consistent estimator with smaller MSE in finite samples. Journal of the American Statistical Association, 1974, 69, 990-996.

Fuller, W. A. Properties of estimators in the errors in variables model. Presented at the 1971 Annual Meeting of the Economic Society, 1971.

Griliches, Z. Errors in variables and other unobservables. Economitrica, 1974, 42, 971-998.

Karni, E., & Weissman, I. A consistent estimator of the slope in a regression model with errors in the variables. Journal of the American Statistical Association, 1974, 69, 211-213.

Kenny, D. A. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82(3), 345-363.

Lord, F. M. Large sample covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 1960, 55, 307-321.

Madansky, A. The fitting of straight lines when both variables are subject to error. Journal of the American Statistical Association, 1959, 54, 173-205.

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

Overall, J. E., & Woodward, J. A. Nonrandom assignment and the analysis of covariance. Psychological Bulletin, 1977, 84, 588-594.

Porter, A. C. The effects of using fallible covariables on the analysis of covariance. Ph.D. dissertation. University of Wisconsin, Department of Statistics, 1967.

Porter, A. C., & Chibucos, T. R. Selecting analysis strategies. In G. D. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, N.J.: Educational Technology Publications, 1974. Pp. 415-464.

Pravalpruk, K., & Porter, A. C. The effect of multiple fallible covariates in analysis of covariance and two correction methods. Presented at the Annual Meeting of the American Education Research Association, 1974.

Reichardt, C. S. The statistical analysis of data from the nonequivalent control group design. In T. D. Cook and D. T. Campbell, The design and conduct of quasi-experiments and true experiments in field settings. Chicago: Rand McNally, in press.

Stroud, T. W. F. Comparing regressions when measurement error variances are known. Psychometrika, 1974, 39(1), 53-67.

Stroud, T. W. F. Comparing conditional means and variances in regression model with measurement errors of known variances. Journal of the American Statistical Association, 1972, 67, 407-412.

Warren, R. D., White, J. K., & Fuller, W. A. An errors in variables analysis of managerial role performance. Journal of the American Statistical Association, 1974, 69(348), 886-893.

Werts, C. E., & Linn, R. L. Considerations when making inferences within the analysis of covariance model. Educational and Psychological Measurement, 1971, 31, 407-416.

Werts, C. E., & Linn, R. L. Analyzing school effects: ANCOVA with a fallible covariate. Educational and Psychological Measurement, 1971, 31, 95-104.

Zellner, A. Estimation of regression relationships containing unobservable variables. International Economic Review, 1970, 11, 441-454.

15. BAYESIAN INFERENCE

Box, G. E. P., & Tiao, G. C. Bayesian inference in statistical analysis. Reading, Mass.: Addison-Wesley, 1973.

Rubin, D. B. Bayesian inference for causality: The importance of randomization. Proceedings of the American Statistical Association, 1975, 233-239.

Rubin, D. B. Formalizing subjective notions about the effects on nonrespondents in sample surveys. Journal of American Statistical Association, 1977, 72, 538-543.

16. FACTOR ANALYTIC REGRESSION APPROACHES

Boruch, R. G., & Wolins, L. Quasi-experimental design: Further explorations. Proceedings of the Social Statistics Section, Annual Meeting of the American Statistical Association. Washington, D.C.: ASA, 1969. Pp. 327-332.

Creager, J. A., & Boruch, R. F. Orthogonal analysis of linear composite variance. Proceedings, 77th Annual Convention of the APA. Washington, D.C.: APA, 1969. Pp. 113-114.

Joreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969, 34, 183-202.

Joreskog, K. G. Simultaneous factor analysis in several populations. Psychometrika, 1971, 36, 409-426.

Jurs, S. G. Factor analysis structural regression of data from the 1968-69 compensatory education evaluation (Research Report No. 47). Parts 1, 2, 3. Boulder, Colorado: Laboratory of Educational Research, University of Colorado, 1971.

Lawley, D. N., & Maxwell, A. E. Factor analysis as a statistical method. London: Butterworth, 1971.

Lawley, D. N., & Maxwell, A. E. Regression and factor analysis. Biometrika, 1973, 60(2), 331-338.

Warren, R. D., White, J. K., & Fuller, W. A. An errors-in-variables analysis of managerial role performance. Journal of American Statistical Association, 1974, 69, 886-893.

17. STRUCTURAL EQUATIONS

Barnett, V. D. Simultaneous pairwise linear structural relationships. Biometrics, 1969, 129-142.

Blalock, H. M., Jr. (Ed.). Causal models in the social sciences. Chicago: Aldine, 1971.

Blalock, H. M., Jr. Causal inferences in nonexperimental research. Durham, North Carolina: The University of North Carolina Press, 1974.

Duncan, O. D. Introduction to structural equation models. New York: Academic Press, 1975.

Goldberger, A. S. Maximum likelihood estimation of regressions containing unobservable independent variables. International Economic Review, 1972, 13, 1-15.

Goldberger, A. S., & Duncan, E. D. (Eds.). Structural equation models in the social sciences. New York: Seminar Press, 1973.

Hobbs, D. A., Jr. On analyzing the effects of policy interventions: Box-Jenkins vs. structural equation models. In D. R. Heise (Ed.), Sociological Methodology, 1977. San Francisco: Jossey-Bass, 1977.

Joreskog, K. G. Structural equation models in the social sciences: Specification, estimation and testing (Research Report 76-9). Sweden: Uppsala University, Statistics Department, 1976.

Joreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969, 34, 183-202.

Joreskog, K. G. Simultaneous factor analysis in several populations. Psychometrika, 1971, 36, 409-426.

Joreskog, K. G., & Sörbom, D. Some models and estimation methods for analysis of longitudinal data. In D. J. Aigner and A. S. Goldberger (Eds.), Latent variables in socioeconomic models. Amsterdam: North Holland, 1976.

Joreskog, K. G. Structural equation models in the social sciences: Specification, estimation, and testing. In P. R. Krishnaiah (Ed.), Applications of statistics. Amsterdam: North Holland, 1977, pp. 265-288.

Joreskog, K. G., & Sörbom, D. Advances in factor analysis and structural equation models. Cambridge, Mass.: Abt, 1979.

Linn, R. L., & Werts, C. E. Analysis implications of the choice of a structural model in the nonequivalent control group design. Psychological Bulletin, 1977, 84, 229-234.

Werts, C. E., Linn, R. L., & Joreskog, K. G. Quantifying unmeasured variables. In H. M. Blalock (Ed.), Measurement in the social sciences. Chicago: Aldine, 1974.

Wiley, D. E. Approximations to ceteris paribus: Data adjustment in educational research. Paper presented at the International Research Seminar on Learning and the Educational Process, Munich, Germany, 1971.

18. RANDOMIZED FIELD EXPERIMENTS FOR ESTIMATING THE EFFECTS OF NEW PROJECTS, PROGRAMS, POLICIES

Ackland, H. Are randomized experiments the cadillacs of design? Policy Analysis, 1979, 5, 223-240.

Baker, S. H., & Rodriguez, O. Random time quota selection: An alternative to random selection in experimental evaluation. In L. Sechrest, et al. (Ed.), Evaluation Studies Review Annual, 1979, 4, 185-196.

Bennett, C. A., & Lumsdaine, A. A. (Eds.). Evaluation and experimentation: Some critical issues in assessing social programs. New York: Academic Press, 1975.

Bernstein, I. N. (Ed.). Validity issues in evaluative research. Beverly Hills: Sage, 1976.

Bock, R. D. Multivariate statistical methods in research. New York: McGraw-Hill, 1975.

Boruch, R. F., McSweeny, A. J., & Soderstrom, E. J. Randomized field experiments for program planning, development and evaluation. Evaluation Quarterly, 1978, 2, 655-95.

Boruch, R. F. Coupling randomized experiments and approximations to experiments in social program evaluation. Sociological Methods and Research, 1975, 4(1), 31-53.

Boruch, R. F. Contentions about randomized experiments for program evaluation. In H. W. Riecken & R. F. Boruch (Eds.), Proceedings of the SSRC Conference on Social Experiments. Boulder, Co.: Westview Press, 1975.

Boruch, R. F., & Riecken, H. W. (Eds.). Experimental testing of public policy: Proceedings of the 1974 Social Science Research Council Conference on Social Experiments. Boulder, Co.: Westview Press, 1975.

Boruch, R. F., & Gomez, H. Sensitivity, bias, and theory in impact evaluations. Professional Psychology, 1977, 8, 411-434.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963. (Also published as Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1966.)

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

Dobson, D., & Cook, T. J. Implementing random assignment: A computer based approach in a field experimental setting. Evaluation Quarterly, 1979, 3, 472-489.

Fairweather, G. W. Methods for experimental innovation. New York: John Wiley & Sons, 1967.

Fedorov, V. V. Theory of optimal experiments. London: Academic, 1972.

Fisher, R. A. Design of experiments. New York: Hafner, 1935.

Gilbert, J. P., Light, R. J., & Mosteller, F. Assessing social innovations: An empirical base for policy. In A. Lumsdaine and C. A. Bennett (Eds.), Central Issues in Social Program Evaluation. New York: Academic Press, 1975.

Glaser, D. Routinizing evaluation: Getting feedback on effectiveness of crime and delinquency programs. Rockville, Maryland: National Institute of Mental Health, 1973.

Lyall, K. C. Some observations on design issues in large-scale social experiments. Sociological Methods and Research, 1975, 4, 54-77.

Mosteller, F. Rerandomization. In D. L. Sills (Ed.), International Encyclopedia of the Social Sciences. New York: Macmillan and the Free Press, 1968.

Riecken, H. W., Boruch, R. F., Campbell, D. T., Glennan, T. K., Pratt, J., Rees, A., & Williams, W. Social experimentation: A method for planning and evaluating social programs. New York, N.Y.: Academic Press, 1974.

Riecken, H. W., & Boruch, R. F Social experiments. Annual Review of Sociology, 1978, 4, 511-532.

Rivlin, A. Systematic thinking for social action. Washington, D.C.: Brookings Institution, 1971.

Roos, N. P. Contrasting social experimentation with retrospective evaluation: A health care perspective. Public Policy, 23, 1975, 241-257.

Roos, L. L., Roos, N. P., & McKinley, B. Implementing randomization. Policy Sciences, 1977, 3(4), 548-559.

Rubin, D. B. Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 1978, 6(1), 34-58.

Suchman, E. A. Evaluation Research. New York: Russell Sage Foundation, 1967.

Weiss, R. S., & Rein, M. The evaluation of broad-aim programs: Experimental design, its difficulties, and an alternative. Administrative Science Quarterly, 1970, 15, 97-109.

Weiss, C. H. Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Weiss, R. S., & Rein, M. The evaluation of broad-aim programs: Difficulties in experimental design and an alternative. In C. H. Weiss (Ed.), Evaluating action programs: Readings in social action and education. Boston: Allyn and Bacon, 1972, pp. 236-249.

19. REGRESSION DISCONTINUITY

Borich, G. D. Interactions among groups regressions: Testing homogeneity of group regressions and plotting regions of significance. Educational and Psychological Measurements, 1971, 31, 251-253.

Boruch, R. F. Headstart evaluation: Use of regression-discontinuity and randomized experiments. Evaluation Research Report NIE-05a/memo, Evanston, Illinois: Psychology Department, Northwestern University, 1974.

Boruch, R. F. Regression-discontinuity designs revisited. Evaluation Research Report NIE-05b/memo, Evanston, Illinois: Psychology Department, Northwestern University, 1974. (Presented at the Annual Meetings of the American Educational Research Association, Chicago, May 1974.)

DeGracie, J. S., & Boruch, R. F. Regression-discontinuity analysis of the impact of Title I-supported reading programs. Evaluation Research Report NIE-10, Evanston: Psychology Department, Northwestern University, 1975.

Farley, J. V., & Hinich, M. J. A test for a shifting slope coefficient in a linear model. Journal of the American Statistical Association, 1970, 65, 1320-1329.

Gujarti, D. Use of dummy variables in testing for equality between sets of regression coefficients. American Statistician, 1970, 24, 50-52.

Gulliksen, H., & Wilks, S. S. Regression tests for several samples. Psychometrika, 1950, 15, 91-114.

Pearson, E. S. Appendix to paper by B. H. Wilsdon, Journal of the Royal Statistical Society, Supplement 1, 1934, 178-192.

Quandt, R. E. Tests of the hypothesis that a linear regression system obeys two separate regimes. Journal of the American Statistical Association, 1971, 55, 324-330.

Quandt, R. E. New approach to estimating switching regressions. Journal of the American Statistical Association, 1972, 67, 306-310.

Riecken, H. W., et al. The priority score allocation design. In M. Guttentag (Ed.), Evaluation Studies Review Annual, 1977, 2, 114-124.

Rubin, D. B. Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics, 1977, 2(1), 1-26.

Smith, H. Fairfield. Relationships among characteristics observed in area clearance tests. Biometrics, 1951, 7, 185-199.

Snedecor, G. W., & Cochran, W. G. Statistical methods. Ames: Iowa State, 1967, p. 432-436.

Thistlethwaite, D. R., & Campbell, D. T. Regression discontinuity analysis: An alternative to the ex-post facto experiment. Journal of Educational Psychology, 1960, 51, 309-317.

Welch, B. L. Some problems in the analysis of regression of k samples of two variables. Biometrika, 1935, 27, 145-160.

Wilson, J. W., & Carry, L. R. Homogeneity of regression--its rationale, computation, and use. American Educational Research Journal, 1969, 6, 80-89.

20. TIME-SERIES AND LONGITUDINAL DATA ANALYSIS

Anderson, T. W. The statistical analysis of time series. New York: Wiley, 1971.

Bock, R. D. Multivariate statistical methods in behavioral research. New York: McGraw-Hill, 1975.

Box, G. E. P., & Jenkins, G. M. Time series analysis: Forecasting and control. San Francisco: Holden Day, 1970.

Box, G. E. P., & Tiao, G. C. A change in level of a non-stationary time series. Biometrika, 1965, 52, 181-192.

Box, G. E. P., and Tiao, G. C. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association, 1975, 70, 70-92.

Bryk, A. S., & Weisberg, H. I. The implications of nonrandom assignment in comparative studies involving growth systems. Proceedings of the American Statistical Association, Social Statistics Section. Washington, D.C.

Bryk, A. S., Weisberg, H. I. Use of nonequivalent control group design when subjects are growing. Psychological Bulletin, 1977, in press.

Campbell, D. T. Focal local indicators for social program evaluation. In M. Guttentag (Ed.), Evaluation Studies Review Annual, 1977, 2, 125-145.

Glass, G. V., Willson, V. L., & Gottman, J. M. Design and analysis of time-series experiments. Boulder: University of Colorado, Laboratory of Educational Research, 1972.

Gottman, J. M., McFall, R. M., & Barnett, J. T. Design and analysis of research using time series. Psychological Bulletin, 1969, 72, 299-306.

Hannan, M. T., & Young, A. A. Estimation in panel models: Results on pooling cross-sections and time series. In D. R. Heise (Ed.), Sociological Methodology, 1977. San Francisco: Jossey-Bass, 1977.

Hibbs, D. A. Problems of statistical estimation and causal inference in time-series regression models. In H. L. Costner (Ed.), Sociological Methodology, 1973-1974. San Francisco: Jossey-Bass, 1974. Pp. 252-308.

Hobbs, D. A., Jr. On analyzing the effects of policy interventions: Box-Jenkins vs. structural equation models. In D. R. Heise (Ed.), Sociological Methodology, 1977. San Francisco: Jossey-Bass, 1977.

Kepka, E. J. Model representation and the threat of instability in the interrupted time series quasi-experiment. Ph.D. dissertation. Northwestern University, 1972.

Nelson, C. W. Applied time-series analysis. San Francisco: Holden-Day, 1973.

Nesselroade, J. R., and Baltes, P. B. (Eds.) Longitudinal research in the study of behavior and development. New York and London: Academic Press, 1979.

Rao, M. M. Covariance analysis of non-stationary time series. In P. R. Krishaiah (Ed.), Developments in statistics. New York and London: Academic, 1978.

Rogosa, D. Causal models in longitudinal research: Rationale, formulation and interpretation. In J. R. Nesselroade and P. B. Baltes (Eds.), Longitudinal research in human development: Design and analysis. New York: Academic Press, 1978.

21. TIME-SERIES APPLICATIONS

Baldus, D. C. Welfare as a loan: An empirical study of the recovery of public assistance payments in the United States. Stanford Law Review, 1973, 25, 123-250.

Box, G. E. P., & Tiao, G. C. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association, 1975, 70, 70-92.

Campbell, D. T., & Ross, H. L. The Connecticut crackdown on speeding: Time series data in quasi-experimental analysis. Law and Society Review, 1968, 3, 33-53.

Caporaso, J. A., & Pelowski, A. L. Economic and political integration in Europe: A time-series quasi-experimental analysis. American Political Science Review, 1971, 65, 418-433.

Glass, G. V. Analysis of data on the Connecticut speeding crackdown as a time series quasi-experiment. Law and Society Review, 1968, 3, 55-76.

Glass, G. V., Tiao, G. C., & McGuire, T. O. Analysis of data on the 1900 revision of German divorce laws as a time series quasi-experiment. Law and Society Review, 1971, 4, 539-562.

Gottman, J. M., & McFall, R. M. Self-monitoring effects in a program for potential high school dropouts: A time series analysis. Journal of Consulting and Clinical Psychology, 1972, 39, 273-281.

Hibbs, D. A. Economic interest and the politics of macroeconomic policy. Unpublished manuscript. Cambridge, Massachusetts: Center for International Studies, Massachusetts Institute of Technology, 1975.

Peleg, I. International politics as a process of value exchange and conversion: Analytical framework and empirical application. Ph.D. dissertation. Northwestern University (Evanston), 1974.

Ross, H. L. Law, science, and accidents: The British Road Safety Act of 1967. Journal of Legal Studies, 1973, 2(1), 1-75.

Ross, H. L. The Scandinavian myth: The effectiveness of drinking and driving legislation in Sweden and Norway. Evaluation Studies Review Annual, 1976, 1, 578-604.

Ross, H. L., Campbell, D. T., & Glass, G. V. Determining the social effects of a legal reform: The British "breathalyzer" crackdown of 1967. American Behavioral Scientist, 1970, 13, 493-509.

Whittaker, G. F. An economic and statistical evaluation of the performance of hospitals in a merged system: An application of a quasi-experimental design. Ph.D. dissertation. Northwestern University, 1974.

Zimring, F. E. Firearms and federal law: The Gun Control Act of 1968. Journal of Legal Studies, 1975, 4(1), 133-198.

22. EX POST FACTO DESIGNS

Scriven, M. Maximizing the power of causal investigations: The modus operandi method. In G. V. Glass (Ed.), Evaluation Studies Review Annual, 1976, 1, 101-118.

Wilson, S. Explorations on the usefulness of case study evaluations. Evaluation Quarterly, 1979, 3, 446-459.

23. COUNT DATA

Bishop, Y., Feinberg, S. E., & Holland, P. W. Discrete multivariate analysis. Cambridge, Ma.: M. I. T. Press, 1975.

Bross, I. D. Misclassification in 2x2 tables. Biometrics, 1954, 10, 478-486.

Goodman, L. A. Analyzing qualitative/categorical data. Cambridge, Mass.: Abt, 1978.

Haberman, S. J. Analysis of qualitative data. Vol. 1. (Introductory topics). New York and London: Academic, 1978.

Krotki, K. (Ed.). Developments in dual system estimation of population size and growth. Alberta Canada: University of Alberta Press, 1978.

24. PATTERN ANALYSIS

Light, L. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcomes. American Psychologist, 1977, pp. 752-759.

Tukey, J. W. Exploratory data analysis. Reading, Mass.: Addison-Wesley, 1977.

25. SECONDARY ANALYSIS

Boruch, R. F. Record linkage in longitudinal and correlational research: Its justification and implications for individual privacy. In T. Dalenius and A. Klevmarken (Eds.), Privacy protection and the need for data in the social sciences. Stockholm: Social Science Research Council, 1976.

Boruch, R. F. (Ed.) New directions for program evaluation: Secondary analysis. San Francisco: Jossey-Bass, 1978.

Boruch, R. F., and Wortman, P. M. An illustrative project on secondary analyses. New Directions for Program Evaluation, 1978, 4, 89-110 (San Francisco: Jossey-Bass).

Bryant, F. B., & Wortman, P. M. Secondary analysis: The case for data archives. American Psychologist, 1978, in press.

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

Dollar, C. M., & Ambacher, B. I. The national archives and secondary analysis. In R. F. Boruch (Ed.), New directions: Secondary analysis. San Francisco: Jossey-Bass, 1978, pp. 1-7.

Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5(10), 3-8.

Hedrick, T. E., Boruch, R. F., & Ross, J. Policy and regulation for ensuring the availability of evaluative data for secondary analysis. Policy Sciences, 1978, 9, 259-280.

Hyman, H. H. Secondary analysis of survey samples. New York: Wiley, 1972.

Light, R. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Schoenfeldt, L. F. Data archives as resources for instruction, research, and policy planning. American Psychologist, 1970, 25, 609-616.

Stockmayer, W. H. Data evaluation: A critical activity. Science, 1978, 201.

Wolins, L. Responsibility for raw data. American Psychologist, 1962, 17, 657-658.

9b

26. META-ANALYSIS

Cook, T. D., & Gruder, C. L. Metaevaluation research. Evaluation Quarterly, 1978, 2, 5-51.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcomes. American Psychologist, 1977, pp. 752-759.

27. THE ROLE OF THEORY AND MODELS OF PROGRAMS AND ITS RELATION TO EVALUATION

Amaro, N. Comments on "Family composition and structure in relation to nutritional and health programs." In R. E. Klein et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 218-225.

Chernichovsky, D. The economic theory of the household and impact of measurement of nutrition and related health programs. In R. E. Klein, et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 227-267.

Daly, J., Davis, J. H., and Robertson, R. Determinants of health and nutritional status. In R. E. Klein, M. S. Read, H. W. Riecken, J. A. Brown, A. Pradilla, and C. H. Daza (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 7-37.

Scrimshaw, S. C. M., and Pelto, G. H. Family composition and structure in relation to nutrition and health programs. In R. E. Klein et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 183-218.

92

28. ORGANIZATION OF IMPACT EVALUATIONS: MANAGEMENT, POLITICS, AND USE OF RESULTS

Chavez, A. Comments on "Political and organizational issues in assessing health and nutrition interventions." In R. E. Klein et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 332-338.

Riecken, H. W. and Boruch, R. F. (Eds.) Social experimentation: A method for planning and evaluating social programs. New York: Academic, 1974.

Stromberg, J. S. Some suggestions for improving applicability and utilization of evaluation assessments. In R. E. Klein et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 339-361.

Ugalde, A., and Enrey, R. Political and organizational issues in assessing health and nutrition interventions. In R. E. Klein et al (Eds.) Evaluating the impact of nutrition and health programs. New York: Plenum, 1979, pp. 309-331.