

SOME RECENT U.S. EXPERIENCES WITH EVALUATION RESEARCH,  
AND THEIR POSSIBLE IMPLICATIONS FOR LATIN AMERICA

Thomas D. Cook,<sup>1</sup>  
Emile G. McAnany,<sup>2</sup>

Paper prepared for the  
CONFERENCE ON THE MEASUREMENT OF THE IMPACT OF NUTRITION  
AND RELATED HEALTH PROGRAMS IN LATIN AMERICA

1-4 August 1977

Panama City, Panama

*BEST AVAILABLE COPY*

- 1- Northwestern University  
Evanston, Illinois
- 2- Stanford University  
Stanford, California

## INTRODUCTION

This paper deals with recent developments in evaluation research in the U.S.A. and with their possible relevance to Latin American evaluation practices. We shall concern ourselves for the most part with empirical summative evaluations. These require the collection of observations within some kind of a planned experimental or quasi-experimental design framework in order to summarize the extent to which an intervention has reached its goals, met the claims publicly made for it, or has led to unexpected effects. Summative evaluations are often distinguished from formative evaluations, the latter being attempts by project personnel to develop diagnostic feedback about the project's functioning in order to make changes in how the project is administered. Summative and formative evaluations differ in (1) purpose (summarizing a project's impact versus suggesting changes in some parts of it); (2) scope (summative evaluations involve a greater concern for generalizability and the quality of causal inferences); and (3) origin (most formative research is conducted by in-house staff as opposed to outside evaluators). Though our major emphasis will be on summative work, we shall briefly discuss the conditions under formative research is desirable and the utility and feasibility of having summative evaluations done by outsiders rather than project personnel.

We have chosen to focus on nine issues that have recently emerged in the U.S. literature on evaluation research. Our purpose in isolating these is more to stimulate discussion than provide answers to any problems. The latter can seem rather glib and presumptuous in light of particular situations at particular sites; the more so when these sites are outside of the U.S.A.

The issues to be discussed are:

1. When is it advisable to evaluate and not to evaluate?
2. Whose questions are to be asked in the evaluations?
3. Which organizational factors are conducive to bias?
4. How feasible is randomization?
5. How desirable are the alternatives to randomization?
6. How do we obtain high quality measurement?
7. How can the treatment and the sampling-measurement framework be monitored so that problems can be detected earlier and practical fallback positions adopted?
8. How can generalizability be increased?
9. Which kinds of benefit cost analysis is it feasible to perform in the human services area?

These questions all deal with the use of evaluation research to throw light on whether specific interventions (or "treatments") cause effects that are presumed to solve a defined problem. The questions are less relevant to two related issues. The first concerns the use of research to describe and evaluate the magnitude of a particular problem so that it might be put on a national or international agenda. The second concerns the value premises that underly specific interventions or specific classes of intervention. Many critics argue, for instance, that health problems in Latin America are fundamentally political and stem from the economic and social structure of the countries. Designing interventions that are limited to health, these critics argue, fails to attack the "real" problem of health. As crucial as these value questions are, we shall not address them here. Instead, we will restrict the discussion to summative evaluations for assessing the impact of projects, irrespective of the philosophy or politics that went into designing the projects.

The potential scope of summative evaluation is perhaps best exemplified by a modified version of Suchman's (1967) typology of evaluation questions. These are outlined in Table 1, together with the means required to answer each question. Suchman's first category -- effort -- has to do with how many persons receive a treatment and who they are. A host of subquestions are involved here, having to do with the availability of services, outreach, initial usage of the treatment, and subsequent usage. Surveys are the usual means of answering these questions, except in the rare case where one is lucky enough to have access to an extensive archive that is kept for auditing purposes.

Suchman's second category has to do with whether statistically significant effects of the treatment can be detected at, say, the .05 level. The means for generating an answer are two-fold, depending on disciplinary traditions. Most educators, psychologists, and medical researchers lean heavily

towards the analysis of data from randomized experiments or quasi-experiments, whereas many sociologists and economists have preferred econometric analysis of cross-sectional data. We shall have more to say about this later.

The third category has to do with whether any observed statistical differences might make a practical difference. Here the task is to determine whether the pattern of magnitude and temporal persistence of effects implies an impact on the human need that led to mounting the intervention in the first place. The difficulties here are obtaining widespread agreement that a particular magnitude or duration of effect is enough to make a practical difference, and obtaining agreement that impact -- as opposed to outcome -- questions should even be asked. Some persons contend that impact questions involve higher standards than questions about simple outcomes and that it is often unrealistic to expect a project to make a socially significant difference ("impact") as opposed to a statistically significant difference ("outcome").

The fourth question relates to social and psychological processes that might mediate or impede effects. A crucial subquestion here concerns the extent to which a promised treatment is actually delivered and an analysis is usually required of the reasons why delivery might be different from what was expected. But other questions about mediators are involved. Observational or interview studies are used to measure potential mediators during the course of a study. Where resources do not permit this, an interview or questionnaire is often used at the end.

The final general question has to do with costs. Usually three subquestions are involved here: the dollar cost per time unit to reach each of the persons or communities in the evaluation; the relative effectiveness of different ways of allocating the dollar costs; and computing a ratio of benefits to costs. These three questions are not equally easy to answer and we have listed them in increasing order of difficulty. The issue thus becomes: Under which conditions is it feasible to get which kind of information about cost and benefit? Clearly, obtaining answers to these types of questions lies within the scope of economists and accountants.

It is desirable in any evaluation to have answers to each of Suchman's five general questions. Indeed, one can confidently predict that, when the evaluator presents answers to any subset of these questions, his audience will be curious to know answers to the others. However, the scarcity of resources often makes it difficult to answer all five questions, and the crucial issue then becomes: How should one choose priorities among the questions? In our experience with U.S. health services, effect and cost issues have loomed largest in evaluating therapeutic interventions designed to improve practice, while delivery and cost concerns have loomed largest in evaluating interventions designed to increase the coverage of services. (e.g., Medicare and Medicaid). In general, process and adequacy criteria have played minor roles. We shall later examine whether these U.S. priorities are meaningful for most Latin American health interventions.

Table 1

## Suchman's Typology of Evaluation Questions

General Question	Research Questions	Means
Effort	<ul style="list-style-type: none"> <li>a. How many persons ever receive the treatment?</li> <li>b. How many receive it for how long?</li> <li>c. What are the demographic correlates of availability and usage patterns?</li> </ul>	<p>survey</p> <p>audit</p>
Effect	<ul style="list-style-type: none"> <li>a. Does the treatment have any statistically significant effect?</li> <li>b. On which subgroups is there an effect or a differential effect?</li> </ul>	<p>"experiment" or</p> <p>"nonexperiment"</p>
Adequacy	<ul style="list-style-type: none"> <li>a. To what extent does the magnitude of impact meet the need?</li> <li>b. To what extent does the impact persist over time?</li> </ul>	<p>impact analysis</p> <p>long-term study</p>
Process	<ul style="list-style-type: none"> <li>a. What are the social factors that mediate or impede impact?</li> <li>b. What are the psychological factors that mediate or impede impact?</li> </ul>	<p>questionnaire</p> <p>interview</p> <p>observation</p>
Cost-Benefit	<ul style="list-style-type: none"> <li>a. What does the treatment cost per person per time unit?</li> <li>b. In which ways might the costs be used more effectively?</li> <li>c. What are the financial benefits of the project relative to its costs?</li> </ul>	<p>economic or</p> <p>audit analysis</p>

1. When is it advisable to evaluate and not to evaluate? Some critics have argued that in the U.S.A. evaluation is a tool for slowing down social change, that many policy makers only invoke the need for it when they see that the alternative is introducing some new practice which is ideologically or financially unattractive to them. Evaluation, these critics argue, permits the appearance of sensitivity and action without commitment to either.

The most frequent response to this is to claim that in many instances we do not know whether a planned innovation will be effective. Lacking this knowledge, the argument, goes, we can pour resources into useless programs which give a false appearance of being effective and create a hope that will inevitably be disappointed. In addition, the argument continues, it is difficult to fade out a program once it has been funded on a broad scale, and so the absence of evaluation results may contribute to reducing resources for responsible experimentation with alternatives that might be of some benefit.

Both of these viewpoints have merit, and we need to use them to abstract a general principle which might suggest when the call for evaluation is more likely to be a sincere call to learn rather than a disguised call for inaction. The general principle we have abstracted is this: When it is known from prior studies, or from a data-based strong theory, that an intervention is effective, then generally (but not always) there is little purpose to evaluating the innovation once again. Let us be concrete. In the nutritional domain, we know from years of research which diet supplements will reduce infant mortality and morbidity. Hence, to evaluate one of the known supplements is generally wasteful and slows down change. But there are particular exceptions to this. First, to know that a supplement is effective once ingested tells us nothing about whether particular people will ingest it. Second, to know that the supplement will be ingested, tells us nothing about whether it will be used as a true supplement rather than a substitute. Third, to know that a supplement (e.g., cow's milk) is effective with one group of people in one part of the world does not tell us that it is effective with all groups. Finally, to know that a supplement is effective tells us nothing about whether the resources or political-will exist in a particular country for maintaining any improvement in health and for capitalizing upon any gains in life expectancy or physical strength that the supplement might confer. When there are genuine doubts about any of these issues, then an evaluation might be carried out and should be primarily targeted towards the issue in doubt. When few doubts exist about a treatment's effectiveness, it may not be worthwhile to do the study. But even then an individual should not assume from his own knowledge that an evaluation is not worth the effort. He or she should consult the studies that seem to suggest the inadvisability of evaluation and then discuss them with other knowledgeable persons.

Another circumstance where evaluation is not recommended by some persons is when individuals are in need and resources exist that are thought to meet the need. In this case, the argument goes: "Better provide resources that might be effective and forego the evaluation than let anyone suffer by withholding a resource for evaluation purposes only." Two very different

responses can be made to this. One is to stress the need to hire or train evaluators with a broad and flexible range of methodological tools who can increase the chances of designing studies that permit giving the resources to all and also evaluating effectiveness. (Typically, he would do this by incorporating a base line other than a local no-treatment control group -- e.g., designs with two different treatments, or nonequivalent dependent variables, or time-series, or retrospective pretests, etc.) The second and more philosophic, response is to say: "While resources may be sufficient for everyone in a particular community, this does not mean they are sufficient for all the persons in need in a particular country. Would one not be doing a disservice to these other persons by implementing the treatment on a wide-spread basis that precluded ascertaining its effectiveness?" In this context, imagine the gain caused by knowing from no-treatment control groups that polio vaccines were effective. The issue here is an old one, having to do with whether it can ever be justified to let some people suffer for the sake of others. In any event, the ethical issue of withholding potentially ameliorative treatments can very often be side-stepped by careful research design, so that one ought always to question the position that one should not evaluate because evaluation requires withholding resources from some persons in need.

A context where summative evaluation is definitely not advisable is when a project has not been in the field for long enough that it can have a fair trial in the evaluation. All too often in the U.S.A. evaluations are conducted on new projects while they are still learning their mistakes and experiencing all the unexpected difficulties that inevitably accompany new projects. The time when it is appropriate to begin summative evaluation depends on many factors, most of which are project-related, and so it is difficult to give a numerical estimate of when testing might commence. But it is clear that evaluation presupposes a considerable project development period.

When, then, should one evaluate, particularly in Latin America where the resources for evaluation are scarce and the evaluable projects have to be chosen with care? Our tentative suggestion is this: "After making sure that the project is no longer in an initial development phase, ask whether it is likely to be implemented on a wider scale if it is successful." Some projects, it will soon be realized, cannot be implemented because they suppose an infrastructure of, say, medical facilities and personnel that are simply not available, or because they involve political changes at a local, regional, or national level that it is unrealistic to expect. To evaluate such projects will do little for the particular country in the short term. We want to stress these last two qualifications, for evaluation results obtained in one circumstance or country need not be used solely in that circumstance or country. Many results have a certain (risky) transportability. Moreover, results that are obtained at one time may have a quite different relevance in the same country at a later date when social and political conditions might be considerable different. Thus, a "narrow-minded" conceptualization of "research utilization" urges one to evaluate only projects that neatly fit the resources and social climate of a particular country; while a "broader" conceptualization of utilization decries the geographic and temporal parochialism inherent in the position we have advocated.

2. Whose Questions Should Be Answered? Evaluations are political. At the broadest level, politics is probably related to the national stand towards evaluation, since evaluation is implicitly founded on value premises concerned with pragmatism and a gradualist approach to change within the particular social context chosen by a particular country. At a lower level, politics enters evaluation in the research questions that are addressed. For instance, in the U.S.A. most persons considered it appropriate to ask whether "Sesame Street" taught economically disadvantaged children. However, not everyone considered it appropriate to ask whether the advantaged learned more than the disadvantaged on the average, because their concern was that the program might be widening gaps and might feed into some persons' belief that "Sesame Street" was therefore part of the national educational problem and not part of the solution.

One faces the same dilemma in the health area in Latin America: Should one only assess the impact of a health intervention, or should one also ask about the differential consequences that might result because of the social groups which do and do not receive the new treatment of which do or do not benefit from it once they have received it. While Suchman's including of "effort" in his typology leaves his position in little doubt, we want to stress that some persons are unwilling to test even the possibility of whether the net benefit of a service is greater for the relatively more advantaged than the disadvantaged. This issue appears to be neatly side-stepped if an intervention is targeted exclusively at poor persons, since then the comparative aspects of the intervention have a low profile. But even here the issue arises because it may be -- as Rogers (1975) has suggested for agricultural development -- that the richer among the poor benefit most and, after a few years, their new wealth helps exacerbate feelings about local inequities. The comparative issue has to be faced head-on when one asks about the impact of, say, nursing services, since they may be made more available to persons who need them less -- either because of where the nurses choose to live, or because they work in hospitals to which only certain persons have access.

It would be wrong to think that the only questions with political/value overtones are distributional. All groups interested in agricultural development projects will ask evaluators to examine crop yields and the like. But not all of them will ask the evaluators to examine how an intervention affects land tenancy and attitudes towards landholders, or cooperation between farmers, and self-help behavior. All of these are outcomes that have very different political implications for different constituencies.

The crucial implication of the foregoing is this: Who gets to ask the major evaluation questions and, hence, whose interests are most directly fostered by the work? The concern seems to be growing in the U.S.A. that past questions have come either from evaluators' guesses about decision-makers' information needs or from evaluators' interpretations of formal project goals, and that little cognizance has been taken of what representatives of other interested constituencies might want to know from an evaluation. For instance in health matters the questions are usually those of the project developers and do not include the questions about a health project that deliverers at

all levels would ask, or that would be asked by host country political figures or representatives of the project recipients. The concern with who asks the evaluation questions is made even more pointed in the U.S.A. by the growing realization of how naive it is to ask only decision-makers' questions. This is because key decision-makers are often not contacted directly and, when they are, their answers are not always analytically precise and unambiguous. Also, the turnover of policy-makers is astonishingly rapid, and events often make a decision-maker's original questions obsolete anyway.

If one considers it important to ask a broad range of research questions reflecting the interests of several constituencies, a technical problem arises. How could one in practice perform a multi-constituency evaluation? Many obstacles to such work exist, most of which follow from the unfortunate possibility that the more questions one asks in social research the lower will usually be the quality of the answers to any question. The evaluator has therefore to find a way to select the more important questions and to create a research design that will answer as many questions as possible. Sometimes, he can do this by specifying new questions in terms of outcome variables to be added to the measurement framework rather than in terms of possible causes that would have to be added to the set of manipulable independent variables.

What does our recognition of the political nature of evaluation mean for research in Latin America? It means we think that everyone should ask of actual and proposed research: (1) Who decides what should and should not be evaluated and are there inadvertent or deliberate biases in these choices -- e.g., are self-help projects evaluated more than centralized ones; -- and (2) Who poses the general research questions, and by what process are they arrived at -- is it the project developer, the sponsor, or the evaluator, and is the appropriateness of the major questions checked with groups that might have different interests in the program? Finally we have to ask (3) Who translates the general questions into specific research hypotheses? -- who decides, for example, whether the general goals of stimulating the achievement of poor children in Mexico means teaching them or narrowing educational gaps? We do not know the answers to these questions for the majority of Latin American evaluations, nor are we sure whether the issue has the same salience there that it does in the U.S.A. But we can all rest assured that sooner or later political leaders, administrators, practitioners, and clients will ask: In whose interests are evaluations in our country being conducted? And this may boil down to: "Now, who decides what should be evaluated?" and, "Now, who decides the actual form of the research questions?"

3. The Organization Factors Conducive to Biased Evaluations. In the U.S.A. three patterns of research organization are suspected as contributing to biased evaluations. The first is where the evaluator works for the organization whose effectiveness is being evaluated, as would happen when the evaluator is a staff member of a community health center. The second is where the evaluator is not on the staff but is hired by the center and reports directly to powerful staff members within it. The third is where the evaluator

is independent of the center but is hired by the very office within the agency that is funding the center. This last relationship may mean that the office has a reputational or budgetary interest in producing positive results. If we extrapolate from these three situations, the preferred organizational structure would seem to be where the evaluator is financially independent of the project being evaluated and is funded (1) by sources other than the organization sponsoring the project or (2) by an office within the sponsoring organization that is not responsible for the project under evaluation.

The form of the links between service providers, evaluators, and funders is clearly the responsibility of the agencies which fund evaluation. Their concern has to be that evaluators will not be deliberately biased or inadvertently coopted. However, it should be noted that the organizational structures preferred on the U.S.A. are means and not ends, and the funder of evaluations should be prepared to conclude that organizational means other than the ones we have described may meet the desired ends. Indeed, he should also be prepared to conclude that the very means that are not preferred in the U.S.A. may sometimes be appropriate there because, for instance, not every evaluator funded from an office that is sponsoring a developmental project will be biased. Bias is only thought to more likely in this case than others, but no one argues that it is inevitable. Decisions about organizational structures which minimize bias can be made on a case by case basis; and the practice that is currently preferred in the U.S.A. is merely a convenient bureaucratic procedure for minimizing the need to decide on a case by case basis.

What does this mean for evaluation in Latin America? For evaluations -- there or anywhere -- that are clearly formative in nature, there is no problem since it is extremely useful if evaluators are members of the project being studied. As for summative evaluations, we are not sure how many can be expected to be truly independent. The reason for this is that there is no evaluation research industry in Latin America -- as there is in the U.S.A. -- and much of the Latin American research seems to be conducted by persons who feel personally committed to alleviating the target problem and often to the philosophy or concept behind the particular project being evaluated. Such "passion" in individuals is by no means "bad", but it has to be balanced against a "dispassion" built into the interpersonal system of research. That is, the potentially committed evaluator has to be closely monitored in critical fashion. Unfortunately, the infrastructure for monitoring rarely exists in Latin America, and this may mean that the potential for bias looms larger here than elsewhere.

How often this potential translates itself into bias is quite a different question. On the one hand, we have been struck by the small number of Latin American studies of which we are aware that found no differences. But on the other hand, there are cases of Latin American investigators who in later studies reversed their own initially favorable reports (e.g., Diaz-Guerrero, Witke, Reyes-Lagunes, and Holtzman, 1976; and Graham, 1976). Our point should not be misunderstood, for it concerns potential bias and not an actuality. However, it would be interesting for the agencies that fund Latin American studies to try to estimate the prevalence of inadvertent

bias, and to detail the practical steps they could take to reduce its prevalence if indeed they came to believe that it is prevalent.

4. How Feasible is Randomization? Most U.S. evaluators believe in the desirability of randomization, but many doubt its feasibility. The skepticism seems to be abating somewhat, largely perhaps because of the accumulating number of projects where random assignment was successfully implemented initially and also successfully maintained for the course of the study. It is important here to be clear about what we are talking about, since randomization for group comparability is often confused with randomization for representativeness. Only the former is at issue here. It has to do with the process of randomly assigning experimental units into different treatment groups in order to create aggregates that are probabilistically comparable to each before the treatment is implemented. Randomization for representativeness has to do with the process of selecting a single group so that the chosen sample is representative of the population from which it was selected within known limits of sampling error. Random assignment for comparability is desirable because it rules out nearly all the threats to internal validity enumerated by Campbell and Stanley (1966). However, it is no panacea, because as Cook and Campbell (1975) have pointed out, random assignment can be associated with other threats to internal validity -- systematic attrition, resentful demoralization of controls, compensatory rivalry, and the like. But these restrictions aside, randomization is the best single procedure we have for facilitating causal inference.

The debate about the feasibility of randomization has, we think, been somewhat side-stepped, albeit temporarily, by altering the question. Now the focus is on conditions which maximize the probability of being able to randomize instead of on whether randomization is or is not feasible. Most scholars believe that randomization is most likely when the local demand for a service exceeds the supply; when several different treatments are to be compared; when individuals (or whatever unit is being used) cannot communicate easily with each other; when units are temporally isolated from each other (as when different groups of people who do not know each other come at regular intervals for, say, job training); and when the persons or authorities granting access to respondents understand the need for randomization and are willing to endorse it. Such conditions have led, in the past, to randomly assigning whole villages to water treatments (Dodd, 1934); intact nursery groups to Plaza Sesamo (Diaz-Guerrero, et al., 1976); individuals to birth control treatments in Taiwan (Freedman and Takeshita, 1969), and individuals to nutritional treatments in Colombia (Sinisterra, McKay and McKay, 1973).

Since spatial isolation is important for preserving treatment groups intact, the unit of assignment is often at a higher level of aggregation than the individual. Villages or neighborhoods are common units. In these cases, financial resources may be strained by having more than a few units in the experiment, and it is common to find, say, only two villages in some experimental group and two in the controls. Then, it is desirable to match the villages on variables that are thought to be most highly related to the outcome variable of greatest concern and then to randomly assign from within

the match. This reduces the likelihood that the two best or worst villages might receive the same treatment. However, it does not guarantee equivalence; it only minimized pretest differences.

Since the cooperation of individuals who can control access to respondents helps in achieving random assignment, it is useful to anticipate their questions. In the U.S.A., most of the questions and doubts of administrators relate to their discomfort at creating focused inequities between people who are deliberately treated differently. Many administrators are loathe to allocate differentially unless there is a clear and socially approved justification for it based on merit, need, seniority, or the like. It is rare to distribute scarce and valued resources by lottery, which is in essence what random assignment is.

The points about randomization to which administrators seem most responsive include the following in which should be made after a brief and lucid explanation of what randomization is: (1) showing him or her that other persons in similar positions have previously permitted randomization -- lists of randomized experiments are particularly useful here (see Boruch and Riecker, 1973, for a list for lesser developed countries); (2) permitting the administrator to authorize access to the treatment to whichever persons or communities he deems sufficiently meritorious or needy, and then to randomly assign from the remainder; (3) assuring the administrator that, should a treatment prove effective, it will then be made available to all the control group members; and (4) showing the administrator that steps have been taken to minimize contact between individuals in different treatment groups and that the study has a low profile. None of these strategies guarantees success; but they are said to increase the chance of it.

5. The Desirability of Alternatives to Random Assignment. The two major empirical alternatives to random assignment are some form of a quasi-experiment or a cross-sectional nonexperiment. The major developer of quasi-experimental designs has been Donald Campbell who now publicly regrets the influence his work has had. His argument is two-fold: first, the quality of causal inferences from most quasi-experiments (interrupted time-series excepted) is lower than we used to believe based on the prevalence with which systematic threats operate and the difficulty of controlling for them statistically; and second, the easy availability of quasi-experimental designs may have caused applied social researchers not to try to implement randomized experiments when they might have been feasible. His position should not be taken to mean a distaste for most quasi-experiments, which are useful if nothing better is available. Rather, his position reflects a concern with the failure to interpret the results of quasi-experiments more critically.

The basic problems with quasi-experiments stem, in Campbell's opinion from the prevalence of selection differences, especially selection-maturation, and from issues of measurement reliability -- statistical regression, unaccounted for variance in covariance analysis, etc. In exacerbated form, these same problems plague cross-sectional nonexperiments of the kind where differences in exposure to a treatment are measured and then correlated with a dependent variable collected at the posttest. (There is no pretest in such designs).

*no ref. given*

Cronbach (1977), an extremely sophisticated quantitative social scientist, has taken issue with Campbell. His position is that applied research is conducted to help make decisions and so it is imperative that the information be available when needed for a decision. If a powerful design is possible within the time-frame, then Cronbach would advocate use of the design; but if it is not, then Cronbach would argue that one should go with whatever is possible -- even if it is a passively correlational cross-sectional study. Cronbach's position implicitly assumes that some information is better than none, whereas Campbell argues that it is possible for conclusions about cause to be dramatically wrong and to be used as part of the rationale for introducing new practices that are in fact harmful and for reducing the scope of practices that are in fact beneficial. In this last regard, Campbell has cited the example of Head Start in the U.S.A. where his fear was that positive results were obscured and distorted by an analysis which made the program look misleadingly harmful. The likelihood of drawing mischievous conclusions is reduced if the results from a weak design are intelligently interpreted and all the limitations and assumptions are listed. It is then up to the potential user to estimate if he wants to use the information despite its highly provisional nature. Unfortunately, results from weak designs are not always wisely interpreted in the U.S.A. today, and by the time results become part of the popular or policy discourse it is often the case that the qualifications to conclusions have dropped out.

The current debate about the desirability of non-randomized alternatives has important implications for Latin American practice, particularly in the health area. It is our impression, first, that there is currently not the awareness either of randomization or of the randomized studies in the third world that have been successfully completed; and second, it is our impression that the unit of assignment is often the community rather than the individual. If so, few units will be in the study and pre-treatment comparability is difficult to achieve. Given these points, it seems to us that small-scale quasi-experiments with communities are more widespread than randomized experiments with individuals.

In this situation, it would be advisable to try to match before random assignment. But if this cannot be achieved, it would be self-defeating not to go ahead and conduct some more imperfect quasi-experiment. However, special care should be taken before analyzing quasi-experiments so that there is full awareness of the limitations of many current textbooks on methodology -- particularly those which suggest that multiple regression procedures adjust for all of the initial differences between nonequivalent groups. And in presenting results there should be a full and public discussion of any limitations to causal inferences and generalizability -- somehow we have to educate policy makers and the general public to tolerate more ambiguous results from social research studies that seems to be the case at present. To help in this, we suggest that Latin American evaluators (or any others for that matter!) should not consider their results as "final" until they have been closely reviewed by knowledgeable persons who try to point out the hidden assumptions and restrictions behind conclusions. We are not sure at present about how often and how closely Latin American evaluations are reviewed. Certainly, they should be reviewed.

6. Measurement Issues. One could write a book about measurement issues in evaluation research, particularly in lesser developed countries. However, we shall focus on only four issues that we consider especially important.

The first revolves around a distinction between proximal and distal measures of outcome, the former being measures close to what is actually delivered and the latter being more remote. For instance, if a nutritional program is targeted towards getting villagers to grow and consume certain vegetables, the most proximal measure would be whether they planted the seeds; a more distal measure would be whether they or their family consumed the resulting crops; an even more distal measure would be whether children became bigger and heavier at a faster than expected rate; and an even yet more distal measure would be whether the children learn more and their life's chances are significantly improved. The problem here is that the probability of an effect is higher the more proximal the measure, while the probability of a socially significant impact is higher the more distal the measure. In addition, it should be noted that project personnel want to be evaluated in terms of proximal measures, since these are more closely related to factors under their control; but policy-makers and others ask about distal measures, since these are the indicators of the social problem which justified the project in the first place. It is clearly advisable to measure both proximal and distal variables, but since resources are finite, where should the stress be?

Our answer, once again, "depends". In particular it depends on the longevity of the project being evaluated. To evaluate new projects by any summative criteria after, say, only a year in the field is premature, and formative feedback is more appropriate. To evaluate new projects by distal criteria after a year is foolhardy, since most projects require time in order to learn their mistakes and solve their initial teething problems. But even established projects should not be evaluated by distal criteria if the interval between pretests and posttests is a year or less. Though there are fewer initial problems in this case, there is the reality that the passage of influence from the proximal to the distal takes place in time.

A second measurement problem concerns side effects. We have begun to learn in the U.S.A. that, for many social interventions, the unintended impacts are every bit as important as the intended ones. Think, for instance, of how the automobile has inadvertently affected the residence pattern of North Americans and consider the consequences of this. Think, also, of research on therapeutic drugs like Thalidomine or the side effects of many pesticides. Since no one can hope to foresee and measure all side effects, the practical question is: How can one increase the probability of measuring and detecting side effects? Often, reading about relevant theory will help, as will frank discussions with persons who have had first-hand experience with projects like the one to be evaluated. But ultimately the detection of side effects depends on a sophisticated on-site monitoring system that is not tied into examining a fixed set of measures. In this respect -- as we point out later -- we have been favorably impressed by most Latin American evaluation research, for extensive on-site monitoring is more common than in the U.S.A.

The third measurement problem we shall discuss relates to unfocused treatments. These are usually treatments aimed, not at alleviating a specific need, but rather at providing a general service. A nutrition project aimed at delivering supplements and enhancing growth would be one kind of treatment, and the provision of nursing services would be another. In the latter instance, the nurse might teach prospective nurses, lecture to community groups, clear up stagnant surface water in which mosquitoes could breed, treat cuts and bruises or diphtheria and leprosy. By which criteria could one evaluate the nurse's services so as to be sensitive to what he has done and so as not to impose criteria on these services that may be appropriate to only a small subset of the tasks that have actually been performed?

One way out of the measurement dilemma inherent in unfocused treatments is to throw up one's hands and say: "Such treatments cannot be evaluated". A second response is to say: "They can only be evaluated in terms of what is delivered and not in terms of overall effectiveness" -- in essence, rather like a formative research project. A third response is to say "They can be evaluated in terms of services delivered and the satisfaction of the people reached" -- an evaluation based on "client satisfaction ratings". A fourth response is to say that unfocused treatment can be evaluated in restricted comparative terms, as when one asks whether nurses are more effective in disease control than, say, paramedics. A final response is to say that, given the passage of enough time, unfocused treatments can be evaluated in terms of the health status of individuals and families they have visited often. In short, there are a host of carefully phrased questions that can be asked, but they all have to be sensitive to the multitude of different tasks represented by unitary-appearing labels -- such as nurse or paramedic.

The final measurement issue we shall mention is response bias. It is usually patently obvious to most persons in treatment groups what the researchers would like to hear. This means that response bias is treatment-related and can masquerade as a treatment effect, the more because respondents in third world countries have considerable reason to want to please researchers who have higher status than them and might be seen as representing formal authorities. Recognition of this problem has led to an advocacy of unobtrusive measures. Trace measures might include the amount of surface water in which mosquitoes could breed, or the number of children observed to have certain symptoms. The problem with such measures is, of course, validity and the often-questionable sensitivity of the measurement. Archival measures are becoming increasingly available all over the world, as communities use indicators to record their progress. Thus in some health projects there may be records from community health centers, while in education there may be school records. Of course, the records all have to be carefully scrutinized before use to assess the possibility of different archiving practices across treatment groups; and it should be firmly recognized that such records are only available for persons who use the organization collecting the records (i.e. the health center or school). Unobtrusive measures have to be used with considerable skepticism and background knowledge, but the search of them should never be overlooked. Nor,

for that matter, should the search for measures that indigenous persons can collect from others. For instance, Ethiopian mothers have been known to record their children's height when asked, and they appear to have done so reasonably validly.

7. The Need for Continuous Monitoring. Time and again in the U.S.A. evaluations have turned out to be technically disappointing, if not an outright waste of the taxpayers' money. The disheartening feature of many instances is that the problems were either predictable or, more likely, they could have been detected early and modifications could have been made so as to keep the evaluation on track or so as to "fallback" to some defensible position. The lesson we have learned is that mechanisms are required for continuously and critically monitoring the evaluation, but in a way that is perceived to be supportive. We are not speaking here of contract monitoring; rather, technical monitoring is at issue. To be sure, there is currently a job category in most federal agencies and foundations called "Technical Monitor", but too few of these persons have the necessary technical background and field experience to be able to detect problems early and solve them. Consultants are often used, of course, but they tend to be used once a problem is visibly serious or at fixed interim stages which do not correspond to crisis points. Advisory boards are also used, but these tend to be composed of luminaries who are so busy that they cannot follow an evaluation in any close detail over time.

On-site monitoring has another crucial function other than the early detection of field problems. All too often the treatment on paper does not correspond very well with the treatment actually delivered to individuals or communities. This has made U.S. evaluators even more aware than before of the need to measure directly how the treatment is delivered and then to use these measures in the data analysis. There is always a potential trap in doing this, for very often more of the treatment is given to those in worst health or social circumstances, and most analyses of such data will inadvertently make the treatment seem harmful. Nonetheless, most observers consider it crucial to measure who receives the treatment (or various parts of it sometimes) for how long.

As we said before, we have been impressed by the care that is put into monitoring treatments in Latin American evaluations, and this may account for the relatively high technical quality and ambition of some of them -- e.g., the Peruvian, Cali, Colombia, and Guatemalan nutrition studies, or the Nicaraguan mathematics experiment. However, we would like to pose a question which implies that the growing demand for monitoring in the USA may not be appropriate to Latin America at this time. The question is: "Do we want at this stage in Latin America to conduct summative, high-cost evaluations of a few projects, or coarser evaluations of more projects?" Each strategy implies different research payoffs. In the first case, one assumes that the project is promising and resources are used to evaluate it as it is; in the other case one assumes that the need is to pick out "successful instances" and so one makes a gross cut from among many projects, knowing that one may miss some positive results that may be small in magnitude but will probably detect most positive results of any magnitude. As we said, this is a question about strategy and we have no ready answers.

8. Issues of Generalizability. There seems to be a growing realization in the U.S.A. that the best formal procedures for ensuring generalizability are the least feasible. Random sampling from a well-designated universe best ensures generalizability; yet how often does one see respondents, settings, times, or measures being randomly selected? Generalizability is next best ensured by sampling multiple instances that are maximally different and then demonstrating that the same cause-effect relationship holds across instances, as would be the case if a nutrition program had similar effects in lowland and highland settings in Guatemala at three different times. Attempts of this kind to extend generalizability are more frequent, but they require considerable resources. The final means of extending generalizability is sampling to obtain "impressionistically modal instances", as when one wants to conduct a study with, say, poor inhabitants of villages and then goes out to find convenient persons and villages that correspond to the target profile. Obviously, this is a weak sampling procedure in that (1) it does not readily permit generalizing to all instances in the target population of villages; and (2) it certainly does not permit generalizing across any other types of settings, persons, times or measures.

The negative relationship between the desirability and feasibility of different means of assessing generalizability should cause any person to hesitate who wants to use the results of a single evaluation to justify wide-scale implementation. Because of this, we detect in the U.S.A. a growing awareness that the proper unit on which to base most decisions about change is the review of several evaluations of a project or program rather than the results of a single evaluation (Think, for example, of the many North American studies of prison rehabilitation, negative income guarantees, and racial integration in the schools). But not any review is useful. Rather, one looks for reviews based on studies which evaluated the same or similar projects in a variety of settings with a variety of different kinds of persons at a variety of times using different measures of what is presumed to be the same outcome construct.

In this respect, it is heartening to note in Latin American research that since apparently successful projects are tested a second or third time in different settings before decisions about wider implementation are made. This strategy is wise. Consider the intensive on-site evaluations we mentioned earlier. It is obvious that the results from them cannot be generalized beyond the experimental settings. But it is more important to note that some of the intensive on-site work, particularly the outreach, is carried out by persons in the development team, many of whom are highly motivated to make the project a success. One has to wonder whether the same initiative and hard work to increase outreach would be manifested if the project became national policy and so became part of an extensive and routinized bureaucratic structure instead of part of a smaller team effort to prove the study a success.

There is no way in any single evaluation that all questions about generalizability will be answered, particularly since one often hears: "Would the same results hold for X ..." where X is not the originally designated

target group. But the evaluator should at least try to discover the target populations of persons, settings, and time to which decision makers want to generalize and he should then sample to approximate these populations as best he can, even if it is only with "impressionistic modal instances".

9. Cost and Benefit Concerns. Neither of the present authors is an economist, and our fleeting knowledge of cost concerns stems from being consumers rather than generators of information. However, we hear growing doubts about the feasibility of realistic benefit-cost studies in many social service areas. The doubts have two bases: First, the difficulty of making a valid point estimate of the causal impact that is not confounded by the other factors; and second, the difficulty of assigning dollar estimates to many of the benefits -- e.g. an achievement gain of n points on some tests; or 40% of the children increasing their weight by 10%, etc. This is not to say that the last criticism concerns all analyses of benefit. It is clearly restricted to cases where the major impact does not have a readily understood dollar significance as it does with manpower training or days spent out of the hospital and at work. But even in these last instances, we should not assume that the money earned or saved exhausts the range of possible benefits. In all benefit analyses, assumptions have to be made, including assumptions about the range of benefits to be considered. These assumptions should be publicly stated and justified.

As for cost-effectiveness analyses, the concern here is with the case where a project has different kinds of outcomes. Consider disease control which reduces the time adults spend away from work, increases the number of surviving children, and reduces the local belief in non-Western medicine. It is not too difficult to discover the most effective means for optimizing anyone of these singly. The problems occur in trying to discover means of optimizing any combination of the three. The problems would be less severe if there were a data-based estimate of the matrix of transitions from one variable to the other. The difficulty, of course, is to compute a matrix that is realistic. Some persons seek to avoid the issue by asking major decision-makers which of the outcomes they value most. Then the most effective means of achieving this are estimated and the other outcomes are ignored. This procedure seems reasonable to us, but one has to remember that not all constituencies with an interest in the evaluation have similar priorities. From our underinformed vantage point, the problem of handling multiple outcomes in cost-effectiveness studies still persists. Making cost estimates by themselves is much easier.

### CONCLUSIONS

We have tried to outline nine problem areas in current North American evaluation research. The list is not exhaustive. It would be presumptuous to contend that the evaluation problems in the United States apply equally to Latin America, and we have not argued that they do. Rather, we have sought to raise questions as to whether they might. If they are relevant, we hope that the attempts we have briefly outlined to improve evaluation may stimulate the thinking of persons who seek to overcome similar problems in Latin America.

## REFERENCES

- Baerth, J.M., Morales, E., Verastegui, G., and Graham G.G. Diet supplementation for entire communities. The American Journal of Clinical Nutrition, 23, 707-715, 1976.
- Boruch, R.F., and Riecken, H.W. Randomized Experiments in Lesser Developed Countries. Final Report: AID, 1973.
- Campbell, D.T. and Stanley, J. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1963.
- Cook, T.D., and Campbell, D.T. The Design and Conduct of Quasi-Experiments for Field Settings. In M.D. Dumette (Ed.). Handbook of Organizational and Industrial Psychology. Skokie, Ill. Rand McNally, 1976.
- Diaz-Guerrero, R., Reyes-Lagunes, I., Wittke, D.B., and Holtzman, W.H. Plaza Seramo in Mexico, An Evaluation. Journal of Communication, 26, 145-154, 1976
- Dodd, S. A Controlled Experiment on Oral Hygiene in Syria. Beirut: Publications of the American University of Beirut, Social Science Series #7, 1934.
- Freedman, R., and Takeshita, J.Y. Family Planning in Taiwan. Princeton, New Jersey: Princeton University Press, 1969.
- McKay, H., Sinisterra, L., McKay, A., Gomez, H., Lloreda, P. Cognitive growth in malnourished Colombian preschoolers. Submitted for publication, 1977.
- Sinisterra, H., McKay, H., and McKay, A. Stimulation of intellectual and social competence in Colombian preschool children affected by multiple deprivations of depressed urban environments. Progress Report #1, University Center for Child Development, Human Ecology Research Station, Universidad del Valle, Cali, Colombia, (November 1971 and September 1973).
- Suchman, E. Evaluative Research. Russell Sage Foundation, New York, 1967.