DOCUMENT RESUME

ED 245 794 PS 014 307

AUTHOR	David, Jane L.
TITLE	Making Evaluations of Follow Through Useful to Decision Makers.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	81
CONTRACT	NIE-P-80-0166
NOTE	31p.
PUB TYPE	Viewpoints (120)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Case Studies; Data Analysis; Data Collection; Early Childhood Education; Educational Experiments; *Intervention; Pilot Projects; *Program Evaluation; *Program Implementation; Research Design; *Research Methodology; *Research Problems; Sampling
IDENTIFIERS	Context Effect; *Multiple Case Study Approach; *Project Follow Through

ABSTRACT

After a brief introduction indicating the value of the first Follow Through evaluation, this document offers ideas about how a new wave of Follow Through approaches should be evaluated. The discussion is based on the premise that the form of an evaluation must be derived from the questions for which answers are sought, the audience(s) to whom the results are directed, what is known about the treatment under investigation, and the size and shape of the treatment. It is asserted that local decisionmakers are the primary audience of interest, that their information needs justify the need to understand the process of implementation and change, and that such contextual information is relevant to federal program designers. Considerations pertinent to the design of a pilot intervention program are addressed. Subsequent discussion indicates limitations of experimental and ethnographic approaches in producing evaluation information and proposes a multiple case study approach as a way of minimizing weaknesses of traditional approaches. Specific attention is given to (1) building a conceptual framework that establishes the topics on which data will be collected and that presents the context in which data should be interpreted; (2) sample selection strategies; (3) the data collection strategy; and (4) stages of data analysis. In conclusion, recommendations are offered for the design of a pilot Follow Through program implementation and evaluation. (RH)

BAY AREA RESEARCH GROUP
385 SHERMAN AVENUE - SUITE 3
PALO ALTO, CA 94306

(415) 326-1067

MAKING EVALUATIONS OF FOLLOW THROUGH USEFUL TO DECISION MAKERS

Jane L. David

This paper was prepared for the Department of Education, National Institute of Education, under Purchase Order NIE-P-80-0166, 1981.

INTRODUCTION

Evaluations conducted under the auspices of the federal government are notorious for their lack of use in decisions concerning the programs under scrutiny. After more than a decade of costly evaluation of Follow Through, there is little evidence that any decisions have rested on these expenditures. The Follow Through program as a whole has been remarkably impervious to change--even in the face of repeated attempts by the Administration to terminate the program. Twelve years after its inception, most of the Follow Through sponsors remain the same and most of the local districts in which they are implemented remain the same. Moreover, there is little evidence of any programmatic influence beyond the Follow Through sites themselves, calling into question the R&D justification.

Although the Follow Through evaluation produced little information of immediate use to decision makers, it was far from useless. In fact, it was probably the single most important experiment on the capacity of evaluation to solve problems. The mammoth effort produced invaluable information to the community of evaluators and researchers--information about the limits of national evaluation for educational decisions. The Follow Through evaluation virtually introduced the notion of implementation of treatment to the world of educational evaluation. It also defined a host of issues from appropriateness of available measures to the realities of maintaining equivalent comparison groups and the limits of post hoc statistical adjustments.

By virtue of defining and drawing attention to these issues, the Follow Through evaluation generated information that was useful--particularly to the research community. But it was useful primarily because it was the first of its kind--not because it was relevant to the questions around which the evaluation was designed. The evaluation did little to answer the questions it originally purported to tackle: Does Follow Through work? Which approach works best? The former was not answered because it was a non-question. Without defining "Follow Through" and "work," the question is meaningless. The evaluation also did not ascertain which approach is best because of the difficulties of comparison groups, measures, goals, etc. that have been well documented elsewhere. This is not meant to suggest that the evaluation was poorly planned or executed. Those who conceived of the planned variations design (albeit as a political move to adjust to drastic budget slashes) took a rational approach in the light of what was then known about defining and measuring educational successes. But the world of educational evaluation and research on educational improvement has changed significantly since the late 1960's. We have learned a lot about the limits of evaluation--much of it from the [illegible] Through experience. We have learned a lot about how schools change (or, perhaps more accurately, about why schools don't change). We are therefore in a much stronger position in designing evaluations for new waves of Follow Through approaches to develop evaluation strategies that have a high potential to produce usable information.

The purpose of the remainder of this paper is to communicate some ideas about how an evaluation of new Follow Through approaches might be designed to maximize the usefulness of the information it generates. I begin with the premise that the form of the evaluation must be derived from the questions one is trying to answer, the audience(s) to whom the results are directed, what we already know that is related to the treatment under investigation, and the size and shape (intensity, duration, number of sites, etc.) of the treatment. I have not been able to obtain specific information on any of these topics beyond what exists in the NIE planning document of October 1, 1980. Therefore I have chosen to make up some characteristics of the proposed first wave of new Follow Through approaches in order to have some concrete examples to draw upon in communicating my views about their evaluation.

Drawing on the ideas contained in the NIE (Shiller et al.) document of October 1, 1980, "Plans for Follow Through Research and Development," I pretend that the first wave of new approaches will consist of three different strategies designed to increase the amount of time devoted to instruction. For purposes of this discussion I define the three strategies (loosely) as follows:

- Strategy A: An intervention that provides an intensive in-service training program for teachers designed to teach specific classroom management techniques that minimize lost time.

- Strategy B: An intervention designed to form a school-site council consisting of school staff and community representatives whose charge is to design ways of increasing the amount of time available for instruction.

- Strategy C: An intervention that provides training to principals in how to reorganize their schools (e.g., via coördination of multiple programs) to protect the part of the school day that is devoted to instruction from interference.

Although these definitions of treatment are vague, they will serve the purpose of providing something concrete to refer to in the following discussion of possible approaches to their evaluation. The discussion will consider the topics given above from which the design of the evaluation must be derived: the audience for the answers, the questions to be answered, and the characteristics of the intervention.

Implicit under each topic is the notion that our prior knowledge suggests what can be achieved and thus should influence the choice of audience, questions, and type of intervention. The topics are intimately interrelated and thus difficult to treat separately. For convenience I discuss first the audience (which subsumes the overriding purpose of the evaluation) followed by the questions and finally implications for the intervention itself. Then I turn to implications for the design of the evaluation.

CONSIDERATIONS

The discussion is intended to support the following claims I make about evaluation:

1. An evaluation cannot be all things to all people; questions and audience must be limited in advance.

2. The evaluation questions should be grounded in what we know from previous research and evaluation--including what is 'answerable' given existing measures.

3. The end result (outcome) of the intervention is of little value without understanding how and why it was or was not achieved.

7

4. The design and implementation of the new approaches must be done in conjunction with the evaluator.

The Audience

Before one can determine what qualities would make an evaluation of the new Follow Through models useful, it is necessary to ask "useful for whom?" One of the biggest problems that has besieged evaluation is that of over-promising by claiming too many purposes. The history of evaluations of ESEA Title I is fraught with illustrations of the problem of trying to serve multiple audiences, each with a different stake in the program and hence in the results of the evaluation, with a single evaluation. Who are the audiences for an evaluation of new Follow Through approaches? Presumably there are potential users at all levels of the educational system: federal, state and local. Within each level, as well, there are potential users with different information needs (that is, with different questions that the evaluation might answer for them). For example, at the federal level, there is an audience in Congress--an audience which probably contains various viewpoints but which pretends to speak somewhat as one in legislation. The Congressional audience views Follow Through as a service program and is thus interested in the question of whether the program is serving intended beneficiaries and the quality of those services. There is also an audience in the Administration--in fact, there are probably multiple audiences in the Administration since there are at least two agencies involved in Follow Through and different agencies usually have different agendas. There may even be more than one audience within the program office itself since there are likely some staff who view Follow Through as a service program and are thus primarily interested in ascertaining whether the program is being administered properly and whether the services are being delivered while others view the primary purpose as research and development and might want to develop and test various new approaches.

The closer one gets to the operating program itself, the more certain questions change from those asked several levels above the program. A district administrator or program director is interested in what programs would 'work' in his/her district or subset of schools. A principal with several Follow Through classes in his/her school is likely interested in those classrooms as a unit--and even in their impact on other parts of the school. A teacher or a parent aide or a classroom specialist of some sort is interested in questions pertaining to the particular classroom. Many parents are interested just in their child and find any assessment of a larger unit not particuarly relevant.

I would urge that the first step in designing an evaluation of the new Follow Through approaches should be to decide upon the primary audience of interest--do not try to meet all the information needs of all the actors. NIE has begun this process and their documents indicate that the primary audience of interest is the local decision maker. I offer strong support for this choice having initially embarked, for this very paper, on an attempt to identify the federal policy issues in Follow Through in order to make inferences about what types of evaluations of the new approaches would be most useful to federal policy makers. I found it exceedingly difficult to identify policy issues beyond those that have existed since the birth of Follow Through and will continue until its demise. Should Follow Through be a service program or a research program? This is not a question for which evaluation can provide reasonable input. This is a question that subsumes issus of equity, of values, and of political support and commitment. Never has the debate between service and research been cast (nor could it be) as a researchable question. "Does Follow Through work?" is similarly distant from evaluation. This is the false evaluation question around which the first evaluation was built and continues to be, at that level, a non-question. Only with much greater specificity does this question become researchable.

Similarly, the question of which approach works best, though often cited as the federal R&D question of interest, is not answerable as our experience with the first round of models has shown. Therefore, NIE's goal of informing local decision makers of promising management strategies seems a reasonable choice of audience.

The Questions

The choice of questions to be answered by an evaluation is intimately related to the choice of audience and our current knowledge about the issues. If local decision makers are to be the primary audience for the evaluation, the design must reflect what is of interest to them, constrained or elaborated by what we as researchers know are the limits of evaluation (e.g., what can be measured well) and factors we know to be important from previous research on school change and improvement. To illustrate this I'll draw on the three hypothetical strategies presented above designed to increase instructional time (A through teacher in-service, B through school-site councils and C through training principals),and assume each is implemented in several schools (say, two in each of three districts). I will take for granted all the arguments that have been presented over the years about the limited usefulness of the planned variations approach. Hence, I assume that determining which approach is "best" according to some predetermined set of outcome measures is not only impossible to implement (because of lack of comparability) but, as I will argue below, not relevant to the needs of any audience.

If it's not a horserace, then what is the comparison of interest? Logically, the comparison of interest might be within a strategy--did the amount of instructional time increase? Suppose, for the sake of argument, that five schools with Strategy A had increases in instructional time while three each with Strategy B and C showed increases. (Assume that, in each, instructional time is measured prior to the manipulation, say spring 1982, and after a year, spring 1983). What information of use does this set of findings communicte? I suggest that virtually nothing would have been learned--not because the results were mixed (which, however, is virtually inevitable), but because what is of use to someone else trying to increase instructional time is WHY and HOW a particular approach worked--not whether it worked. WITHOUT THE HOW AND WHY, IT IS IMPOSSIBLE TO MAKE INFERENCES ABOUT WHETHER IT WOULD WORK IN ANOTHER SETTING. Understanding what factors in the context facilitated or hindered the attempt to increase instructional time is critical information for one trying to implement an approach in a particular context beyond those studied. Outcome measures by themselves are of little value in decisions of this type.

The finding most widely agreed upon and cited from the Follow Through Planned Variations experiment supports this argument; to wit, the finding that there was as much variation between sites within sponsor as between sponsors. In addition to this source, a growing number of studies in recent years have confirmed the notion that the particular features of the context in which a change is being implemented are overwhelmingly associated with the success or failure of the attempted change. (See references).

Returning to the audience of interest, let's presume that the target audience is a district level administrator since the district is the grantee in Follow Through. What does such a person want to know? Typically, thoughtful administrators want to know how Approach X worked in a district (school) similar to theirs. Such an administrator might claim interest in knowing only whether (a) there is an outcome measure that documents success and (b) whether the disrict or school studied has characteristics or circumstances similar to theirs. Most administrators are sophisticated enough to know that what works well in one situation may not work at all in a completely different setting (e.g., one which has a strong union, a recent earthquake or a Spanish speaking population as compared to one that doesn't).

Those who are more sophisticated may recognize that success also depends upon the staff involved, their desire and ability to accomodate change (is this the tenth innovation in five years?), the match between what they are currently doing and the proposed change, and generally what new demands will be placed upon all involved in impelmenting the new approach. This might include the particular educational philosophy of a particular principal or of a group of parents active in improving the schools.

Design of the Intervention

The primary purpose of this paper is to suggest how a new wave of Follow Through approaches should be evaluated. It is beyond the paper's scope to suggest how the interventions should be designed and implemented; however, the evaluation is inextricably linked to the design of the intervention. Therefore, I want to communicate some considerations that should go into the design of the pilot. The first consideration is the role of the evaluator. It is critical for the evaluator to be involved in the design of the intervention. Because the primary purpose of the pilot is to learn about certain interventions, the choice of audience, questions, measures, prior knowledge should all affect the shape of the interventions themselves, not just the evaluation. If NIE designs certain interventions and determines the sites in which they will be implemented and the duration of the intervention, AND if these decisions are made without regard to the eventual audience for the evaluation, questions to be answered, etc., there is little point in having an evaluation. Even if NIE does consider all the above in their design, unless NIE is to conduct the evaluation itself, few evaluators would be happy to step into a situation with so many constraints already imposed.

A second consideration in designing the intervention is that of duration. We are in a rapidly changing world. Given declining enrollments, school closings, reduced funding, population shifts, etc., studying the process of change over a period of years may have little relevance for the world that will then exist. This suggests that a ten year undertaking is not useful (especially given administrative considerations such as changing leadership, evaluation staff, and so on). Moreover, with most Congressional cycles running under three years, it makes little sense to design studies dependent upon federal funding that exceed these cycles (from the initial design to the final report--not just data collection).

In view of these severe time constraints, it makes sense to consider ways of obtaining intermediate results. If the ultimate outcome is increased student achievement, which seems inevitable, it would be valuable to design interventions for which there are intermediate outcomes of interest. For example, interventions designed to increase achievement through increased time devoted to instruction, could report on increases in instructional time prior to measuring achievement.

Finally, as the example of instructional time suggests, the intervention should be shaped by the reality of schools as organizations. Given the importance of context, it makes sense to take that context into consideration in designing interventions rather than shaping the intervention in terms of children without regard to context. Trying to change the context in which children learn (which requires dealing with the school as an organizational entity) rather than changing the children might result in interventions which have a greater chance of producing change. (See Henry Acland, "On Structure").

IMPLICATIONS FOR EVALUATION

In the preceding sections I have suggested that a useful evaluation must be focused on a particular audience and a particular set of questions that are relevant to that audience. In particular, I chose local decision makers as the primary audience of interest and justified the need to understand the process of implementation and change by appealing to their information needs which are context specific. Before moving on to the implications for evaluation of these choices, I submit that this type of understanding is equally relevant to the federal decision maker. Although certain federal decisions require national data (populations or samples), such as questions of prevalence and other numerical information, federal actors in the business of designing and administering programs to improve education need to understand how the context affects the process of implementation and change in order to make federal policy that will be effective. In this section I try to show the limitations of traditional experimental and ethnographic approaches in producing this type of information and propose in their place an approach that minimizes the weaknesses of each approach.

Limitations of Experiments and Case Studies

The presumed advantages of an experimental approach lie in its potential to isolate causal factors and to produce generalizable findings. These advanatges can only be achieved under ideal conditions, e.g., random assignment to treatment and random sampling from a well-defined population. Random assignment is seldom feasible and post hoc compensations are far from satisfactory for isolating causal factors. Random sampling is feasible but requires, for most purposes, that the sample be large and the factors of interest be precisely specified in advance. If we could list the ten factors most likely to explain differences in agreed upon outcomes of interest AND list in question form the fifty factors that reflect our best guesses as to additional explanatory factors, we could define a sample that might result in the desired information--IF we were right in our a priori choice of variables. We could choose the sample to represent variation on the first ten and design survey instruments based on the fifty. If we were wrong, however, we would have learned very little at considerable expense. Unfortunately, the evidence suggests we WOULD be wrong since we are just beginning to learn what factors are important to understand in predicting how humans change their behavior in complex organizational settings.

We have already learned that the factors that we originally expected to be good predictors of educational outcomes--sex, age, and the usual raft of other child and teacher characteristics--are woefully inadequate in predicting the results of attempts to change. From a number of change efforts, we have learned that predictors of change include such factors as how a school is introduced to an innovation, whether school staff were part of the decision to implement a given strategy, and whether the strategy was compatible with ongoing enterprises in the school. These types of explanations suggest that the important correlates of change cannot be captured by simple two and three-way interactions.

On the other hand, traditional ethnographic or case study approaches, while able to yield more rich and relevant information, are of minimal use to federal policy makers concerned with making statements applicable to the nation. Although arguments exist for generalizing from a single case (see Kennedy, "Generalizing From a Single Case Study"), a federal decision maker would have difficulty defending policy based on an evaluation of a single case.

Another limitation of case studies is the quantity and type of information obtained. Ethnographers are loath to constrain data collection in advance by preconceived notions. Yet, if one does field work without a carefully predefined structure that does in fact constrain data collection, the result will be a mammoth amount of undigested information--information that may be irrelevant, untrustworthy, and extremely difficult to decipher. Case study investigators are also loath to draw conclusions, preferring to leave all inferences to the reader--an extremely burdensome task for a decision maker.

But suppose we blend the two approaches--experimental and case study--in a way which preserves the structure and generalizability of the traditional experiment with the richness and relevance of the information gained in a case study approach. I call such an approach the "multiple case study approach" which is actually a shortened form of "multiple site, structured case study approach." (This phrase and the discussion that follows draw heavily on the ideas presented in Greene and David, "Generalizing from Multiple Case Studies," in which the supporting arguments are more fully explicated).

The Multiple Case Study Approach

The multiple case study approach rests on building a conceptual framework at the beginning that lays a map of the territory, followed by careful sample selection done purposefully to insure variation on certain factors of importance. The conceptual framework also serves to structure data collection insuring that the data are comparable so that the analysis, which looks for similarities and differences across cases, can be conducted with integrity.

Conceptual Framework

A carefully developed conceptual framework is the backbone of a multiple case study approach. I hesitate to use this phrase because it conjures up images of the obligatory literature reviews and references to theory seen at the beginning of many research reports and never referred to again. The conceptual framework of which I speak serves a far more pervasive role in the conduct of the study. At the beginning, the conceptual framework serves to organize existing knowledge about school change. In the illustrations given above, the conceptual framework would draw on the general literature of school change as well as research specifically concerned with increasing instructional time.

The conceptual framework serves two critical functions. First, it limits the topics on which data will be collected. Second, it presents the context in which the data should be interpreted. By identifying the important components of the intervention and the environment in which it is implemented, the conceptual framework reduces to a manageable set and makes concrete an otherwise unbounded set of concepts (the whole field of personal and organizational change). From this set are generated the topics on which data will be collected and the classes of appropriate respondents. For example, a topic might be the role of the principal in the school (the principal as an source of explanation for change or lack thereof in teacher behavior). In a particular school, the interview with the principal would be in part determined by the findings about change in the teachers. Hence the questions would be different from one school to the next, depending upon what else was going on in the school.

Through identifying the topics and the context that shapes their meaning, the conceptual framework structures and limits the data collection but does not constrain it a priori to specific questions as do survey instruments.

Moreover, it serves to communicate to readers of the final report the particular viewpoint of the evaluator so that the meaning of the findings can be judged. As such, the conceptual framework also serves as a guide to sample selection and as the basis for site visitor training to insure comparable data. Each of these areas is described below.

Sampling

Careful sample selection is critical in conducting a multiple case study for it is the basis upon which generalizability of the findings will be justified. It must be done purposefully, drawing on elements of the conceptual framework. In structure it is analagous to sample selection in an experiment. Just as we would incorporate into any experimental or quasi-experimental sampling plan those factors anticipated in advance to be powerful explainers, so would we here, but with one important exception. We need to insure variation on those factors expected to explain outcomes, but we don't need to include all combinations of all levels of each factor.

For example, suppose we want to implement the school-site council concept in six schools. We have a pretty good idea that:

a. Teacher support for such a council is an important predictor of its eventual success.

b. It is easier to implement changes in a small school than to a large one.

c. Administrative support is important in instituting a planning and decision making group.

d. Staff stability is important to the creation and maintenance of such a council.

Obviously, each of these factors can be measured in a variety of ways and with varying degrees of confidence. For purposes of sample selection, however, it makes sense to limit the factors to those that can be measured in advance with ease and confidence. Hence, for example, I would eliminate teacher support because it is extremely difficult to measure accurately from a distance. And unless one can develop a good proxy for administrative support that can be measured at long range, I would be equally wary. School size and staff stability, on the other hand, can be measured at a distance with relative ease and accuracy. NIE can use this type of information to select a sample (both for the implementation of new approaches and for their evaluation) according to one of two strategies.

One strategy is to select sites for the pilot that are high on all the factors anticipated to affect the desired changes. This rests on the argument that it is so easy to have NO effect in educational manipulations that it makes most sense to stack the deck in advance as much as possible. This argument says that you will learn more from successes than failures--not only because failure is so common but because there is no way to isolate reasons for failure and therefore to make valid inferences from such cases. Hence, maximizing the chance for success is important for constructive learning. District administrators need to know far more what is likely to facilitate change than what is likely to pose barriers.

For this approach to be useful, it is necessary to demonstrate that the selection procedures are likely to yield the "best" cases. Thus one must justify the choice of selection factors and the choice of sites representing high levels on the factors (as judged, for instance, by a consensus of experts and practitioners). If the selection process is defensible, then this approach maximizes the likelihood of finding successes and in essence provides a test of the hypothesis that the intervention can bring about change. This approach does NOT, however, provide a basis for generalizing about the conditions under which success will occur. In this context, it is an exploratory study; if there are successes, one can speculate about their explanations. A different approach, shown below, is needed to confirm the explanations for change.

The second strategy is to select a sample to achieve maximum variation on the factors of interest (where maximum means representative of the full range of variation in the nation). Suppose that we have chosen school size and staff stability as the selection factors, with a high, medium, and low instance of each factor. We do not need to include all possible combinations but we do need to insure that the sample in which the intervention will be tried contains a high, medium, and low school on each of the two factors.

Since the selection factors usually must be measured at a distance, they will be serving as proxies for other sources of variance that may be either too difficult to measure from a distance, impossible to define precisely, or factors for which a relationship is known to exist but nothing is known about the form of that relationship. Hence, school size may have broad support among researchers as an important concommitant of ability to change without their having much idea of the intervening processes through which school size affects the process of change. With a range of school size in the sample, one may discover, for example, that an important key to implementation is frequent informal communication with the principal and that this is simply easier to accomplish with a faculty of 12 than a faculty of 30.

But what allows us to generalize such a finding to sites not included in the sample? First, we must have confidence in the finding for the sample. This means that we have looked at situations in which the likely explanatory factors vary enough to draw conclusions about which ones are in fact having the greatest effect, and under what limiting conditions. On this basis, we should be able to convince other researchers that our explanation is plausible and that we have considered and rejected all plausible alternatives. These claims are not ultimately provable (in any paradigm)--they are judgments made by knowledgable persons, based on their abilities to think of alternatives, to test them, and to persuade or be persuaded that the findings are valid. (In an experiment, these judgments must be made before data are collected and form the basis for the hypotheses to be tested and the factors on which a sample is selected). Given a valid finding in the sample, the basis for generalizing to other sites rests on evidence that the sample contains variation on plausible explanatory factors representative of the variation that exists in the population of interest.

Data Collection

So far we have developed a conceptual framework that identifies those elements that we deem important in evaluating the impact of new Follow Through approaches. We have also chosen a sample on the basis of those factors in the conceptual framework that are widely agreed to affect the impact and that are measurable from a distance. The next step is to translate our preconceived notions of what is interesting into a data collection strategy. We seek comparability across units in a multiple case study design for the same reason as in an experimental design: to be able to make valid inferences about the relationships between differences in outcomes and differences in conditions.

In any case study data collection, the data collector (or site visitor) himself or herself should be viewed as the data collection instrument. In a multiple case study design, the interview guide provides the structure within which data will be collected. The guide is derived from the conceptual framework and organized by the questions to be answered in the analysis. It is only a guide, however, and as such says little about the necessary amount of detail, or the particular respondents to include or the ways to ask questions, or how to know when enough information has been obtained on a given topic.

Given the context-dependent nature of the data, specfic answers to these issues will vary from site to site--but site-specific guides would foreclose comparability of data across sites. Therefore it is necessary to train the data collectors so that they have a shared understanding of the purposes of the data collection (as well as specific skills for maximizing the validity of the data that are collected). Shared understanding can best be accomplished through involving the data collectors in the development of the conceptual framework and interview guide. A shared veiwpoint and understanding of what constitutes reliable and valid data cannot be accomplished in one-shot training sessions but must evolve from continuous immersion in the concepts and goals of the study. Specific skills for maximizing internal validty can be imparted in a more structured way through formal training and rehearsal in methods such as cross-examination and triangulation. Through simulations of on-site data collection, data collectors can gain experience in probing, in developing multiple approaches, and in drawing inferences on the basis of multiple perspectives.

The particular purposes of each study will dictate how the data are transcribed from field notes. To the extent that an interview guide reflects the categories in which the analyses will be reported, it is useful to write up the field notes in the format outlined by the interview guide. Whether or not this is the case will be a function of how closely the conceptual framework matches the reality found in the sites. In conducting a multiple case study, it is a tremendous advantage to have the luxury of longitudinality--at least to the extent that more than one wave of data collection can be conducted, even if within one shool year. If additional waves of data collection are possible, then the conceptual framework can be revised after each visit to better reflect reality. This can compensate somewhat for omissions in the original conceptual framework or anticiapted relationships that are not supported by the data.

Analysis

The first important stage of analysis occurs while the data collectors are in the field. Each data collector, in the process of gathering information through interviews and observations, has implicitly generated and tested innumerable hypotheses. Choices about whom to interview, what questions to ask, how far to go, etc. are made on the spot, based on the data collector's knowledge, experience and judgment of what is there is to be learned. The data collector is constantly developing hunches about connections between events and choosing questions that will test those hunches. The hunches (or operating hypotheses) are constantly revised, retested, and revised and retesteo again until the data collector has confidence through multiple sources and perspectives that the story is internally consistent and inherently plausible.

The second stage of analysis occurs after a round of data collection. The analyst(s) must first become familiar with each case and draw conclusions from individiual cases so as to connect the features of the local context with the change being studied. Then the analyst(s) conducts pairwise comparisons in which tentative conclusions based on one case are systematically tested against each of the other cases. The purpose of these case-by-case comparisons is to fine tune, modify, and refine the propositions to that they are expressed precisely to reflect the limiting conditions revealed by the patterns of findings across all the cases (e.g., x is true in large schools or y is true in schools with strong principals). If the amount of modification required to make a proposition hold in all instances is excessive--amounting to a completely site dependent phenomenon--the proposition is dropped as uninteresting. The conclusions that remain after this obstacle course of pairwise comparisons are finally presented with illustrations drawn from the cases, in a clear and concise form that can be easily read and understood by the audience of primary concern.

CONCLUSIONS

To maximize the usefulness of what is learned from a new wave of Follow Through approaches, the design of both the pilot and its evaluation should be grounded in reality. The two most salient features of reality are first, that attempts to change educational practice are context dependent and second, that there are limits on what CAN be learned about the effects of an intervention. Therefore, this paper has presented arguments in support of the following recommendations:

-The pilot (including the characteristics of the intervention, its scope and duration) should be designed in conjunction with the evaluation.

-The audience for the evaluation and the questions to be answered should be focused and realistic (both in terms of the types of information useful in decisions and in terms of what we can answer).

-The primary goal should be to understand how and why an intervention works not just the end result. The evaluation should rest on the assumption that the context is something to be examined NOT something to be controlled.

-The proposed multiple case study approach provides a way of maximizing the usefulness of the data collected without sacrificing generalizability.

No experimental approach will result in unambiguous findings; neither will a multiple case study approach. Both approaches rest ultimately on the experience, knowledge and ability of the evaluator(s). There are clearer and more generally accepted rules for conducting experiments than there are for conducting multiple case studies. But these rules, which are inevitably broken in field experiments, are not designed to elicit the types of data that are likely to be used by decision makers trying to improve educational experiences for children. Therefore, it makes sense to move in the direction of refining the type of methodology that is built around increasing our understanding of a complex world and hence more likely to produce information of immediate use.

REFERENCES

Acland, H. "On Structure." Paper commissioned by the National Institute of Education (1981).

Berman, P. and McLaughlin, M.W. "Federal Programs Supporting Educational Change, Volume VIII: Implementing and Sustaining Innovations." The Rand Corporation, R-1589/8-HEW, Santa Monica, CA (May 1978).

Cronbach, L. J. and Associates, Toward Reform of Program Evaluation. San Francisco: Jossey-Bass Publishers (1980).

Elmore, R.F. "Organizational Models of Social Program Implementation." Public Policy, Vol. 26 (1978), 209-217.

Greene, D. and David, J. L. "Generalizing from Multiple Case Studies." In progress.

Kennedy, M. "Generalizing from Single Case Studies." Revised version of paper presented at AERA (1978).

Stearns, M., Greene, D. and David, J. L. "Local Implementation of PL 94-142: First Year Report of a Longitudinal Study." SRI International (1980).

Weatherly, R. A. and Lipsky, M. "Street-Level Bureaucrats and Institutional Innovation: Implementating Special Education Reform." Harvard Educational Review, Vol. 47 (May 1977), 171-197.