DOCUMENT RESUME

ED 244 732 PS 014 292

AUTHOR Kennedy, Mary M.
TITLE Three Versions of Follow Through: An Analysis of Yield.
INSTITUTION Huron Inst., Cambridge, Mass.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE [81]
NOTE 33p.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Costs; *Demonstration Programs; Early Childhood Education; Federal Government; Government Role; *Program Design; *Research Design; *Research Problems
IDENTIFIERS Planned Variation; *Project Follow Through

ABSTRACT

Three possible versions of Project Follow Through are discussed to explore the kinds of knowledge that each approach would yield and to describe alternative ways in which the program could maintain the unique feature of federal coordination of research. Versions compared include (1) a federally coordinated research program designed to generate knowledge about the relative advantages of different combinations of Head Start and Follow Through; (2) the version that actually occurred (the planned variation approach), which contrasted a variety of curricula to determine the nature of effects each produced in disadvantaged children; and (3) small, federally coordinated pilot tests of new, alternative instructional techniques. The comparison is guided by a list of seven research/program dimensions, including nature of programmatic variations, utility of the study, likelihood of measuring relevant outcomes, duration of the study, size of the study, inferential problems, and cost/benefit implications of the above considerations. In conclusion, eight arguments weighing heavily against the third variation are presented, and an alternative third version is offered for consideration. In this latter version, the program would continue as before, and the federal government would support research into all major program dimensions (such as the nature and implications of disadvantage, the organization and delivery of noninstructional services, variations in compliance with federal regulations, program implementation, and characteristics of program models). (RH)

THREE VERSIONS OF FOLLOW THROUGH: AN ANALYSIS OF YIELD

Mary M. Kennedy
The Huron Institute

The Follow Through Program is one of a number of federally supported educational demonstration programs. These demonstration programs as a group have been the subject of a great deal of analysis. For the most part, these demonstration programs can be characterized by certain common characteristics. First, grants are awarded competitively, so that projects must demonstrate not only need but also an innovative idea for meeting that need. This characteristic distinguishes them from formula-grant programs in which projects receive funds based on their number of eligible children or schools. Demonstration programs are also distinguished by their emphasis on demonstrating and disseminating their innovative ideas, whereas formula-grant programs are primarily interested in assuring a minimal amount of services to a specified group. The general idea is that once demonstration projects receive their grant, they will develop and implement their programmatic ideas, evaluate the effects of their ideas on the children or families they serve, and, assuming the evaluation results are positive, they will disseminate their ideas and encourage others to adopt them. Their purposes are often vague, and since in most cases the several projects supported by a demonstration program are quite variable, it is not clear how one can judge the effectiveness of the program as a whole, independent of the effectiveness of its individual projects.

Like other demonstration programs, Follow Through is designed to combine services with research and demonstration, but it is unique in that the research aspects of the program have traditionally been coordinated at the federal level, rather than being left to local discretion. As a result, grants are not awarded to the better ideas; rather, grants are awarded to projects which are willing

to cooperate in a large-scale study. Despite these features, however, the program is similar to other demonstration programs in that its knowledge-production value as a total program is hard to assess.

My task in this paper is to discuss alternative ways in which the Follow Through program could maintain its unique feature of federally-coordinated research, and the kinds of knowledge benefits that each approach would yield. I have selected three particular versions of Follow Through to discuss. The first version is based on the fact that Follow Through was originally designed to strengthen the effects that the Head Start program had on disadvantaged children. If research were designed to learn more about how those gains could be strengthened, it might have been designed to contrast, say, alternative ways of coordinating the two programs. The second version is the one that in fact has predominated in the Follow Through program for over a decade. That version consisted of contrasting a variety of curricula to determine the nature of effects that each one produced in disadvantaged children. This second version has come to be known as the planned variation approach to research, although clearly the first version would also entail planned variations of educational approaches. The third version I will discuss is the one NIE is currently contemplating. Although NIE's ideas are still formative, it appears that they also would entail comparisons of alternative instructional techniques.

In order to discuss these different versions of Follow Through, I will need some means of characterizing them, a strategy for identifying their similarities and differences. There are a variety of characteristics that could be used, and I have selected seven that I think will capture the most important dimensions. These are as follows:

1. The nature of the programmatic variations. Each of these versions of the Follow Through program involves comparisons of some dimension of education--

the duration, organization, substance, or amount of services offered. Of particular interest in assessing the potential value of these studies is whether some of those dimensions are more easily implemented than others, and hence more easily varied for research purposes, whether some are cheaper or more transportable than others, or whether some are of more interest to different research audiences than others, and hence more likely to generate research findings that are considered valuable.

2. The utility of the study. My assumption here is that certain kinds of programmatic variations are more likely to be useful than others, but that the utility of any planned variation study may vary from audience to audience. While an administrator, for example, may be more interested in administrative mechanisms, a teacher may be more interested in classroom practices. Hence, utility depends not only on the dimensions of programs that are varied, but also on the intended audience for the study results.

3. The likelihood of being able to measure the relevant outcomes. There probably is no educational program for which all relevant outcomes can be measured. That may not matter to those who are actually implementing the program because they can see various indicators of these outcomes every day in their work. But it matters if the outcomes are to be aggregated across a variety of projects and communicated to interested audiences. Hence, an important characteristic of these different versions of the program is the extent to which these outcomes can be measured in sufficiently standardized ways that the results can be communicated to others.

4. The duration of the study. Issues arise and decline in importance rather quickly in education. We attend to the needs of disadvantaged children and suddenly realize we have neglected the handicapped. We attend to the needs of the handicapped and suddenly realize our population has declined and our

budgets are suddenly out of line. Studies that take too long to complete may easily address issues which no one has the time, energy, or funds to deal with when the results finally become available. Hence, the probable duration of a study has some bearing on its value.

5. The size of the study. Not all studies must be large but there may be some that must be, and if a study must be large in order to vary the programmatic dimensions of interest in such a way that reasonably interpretable results will be generated, then the costs of the study must be taken into account when considering its yield.

6. Inferential problems. I am not in that camp who insist that all field studies should employ random assignment of children, schools, or projects to programmatic variations, but I am also not willing to give up all aspects of experimental control. Different quasi-experimental designs pose different inferential problems, and one way to judge the probable yield of a study is by the extent to which it will pose serious inferential problems.

7. The cost-benefit implications of the above considerations. Given the nature of the programmatic variations and their utility to various audiences; given the probable outcomes to be measured, and given the kinds of inferences that may or may not be possible, the duration of the study and its size; it should be possible to make some non-quantitative estimates of the benefits of each version of Follow Through, relative to its costs.

Before going into the analysis of each of these versions of Follow Through, I should point out that there are a variety of kinds of yield that I am not discussing, primarily because they are too difficult to estimate in the abstract. Many of these are similar to the yields that demonstration programs in general hope to have. One could, for example, discuss the contributions of any demonstration program to the general state of the art in education. Programs influence education by training teachers who go on and work in other areas and who may

carry the program's influence with them. They influence the researchers whose innovations are tried out, and to the extent that these researchers convey their research findings and their reflections to their colleagues, demonstration programs may contribute substantially to what is considered "conventional wisdom" in a given field. This has certainly occurred in the version of the Follow Through planned variation study that has just been completed. But yields such as that, while valuable, are not possible to estimate in the absence of a real study which has had such a yield, and so will not be considered in my analyses of hypothetical versions of Follow Through. After a short overview of three versions of federally coordinated research that could be carried out in Follow Through, this paper will analyze them in more detail by going through the seven characteristics described above.

The Three Versions

President Johnson chose the field of education to conduct his biggest battles in the war against poverty. He began with the Head Start program, and shortly thereafter decided that disadvantaged children needed more than what Head Start provided. He wanted an early elementary program that would follow through on the start these children had already been given. His vision was of a large program that would serve essentially the same population of children that had been served by Head Start, but when the Congress appropriated too small an amount of money, the program administrators began thinking of the program as a demonstration program, and wondering how it could be designed to serve a knowledge-production function. What they decided was to support the development and implementation of a variety of educational approaches, like most other demonstration programs do, but rather than supporting each approach in only a single project, they decided to implement each approach in multiple projects, and to study their effects over time.

But suppose an accident of fate had put a slightly different group of people

there at that time--one who was more interested in the concept of following through on Head Start. Such a group may have still wanted federal coordination of the research aspects of the program, but would simply have preferred studying Head Start-Follow Through linkages to studying curricular models. The study that this group might have generated will constitute our first version of Follow Through: a federally coordinated research program designed to generate knowledge about the relative advantages of different combinations of Head Start and Follow Through.

The second version of Follow Through is the one that actually occurred. In fact, instead of considering the link to Head Start as a researchable variable, it was held as a programmatic constant, and the administrators studied curriculum variations. They hired a number of different research organizations to develop both the theories and their practical implications, and to assist projects in implementing these approaches. Under prevailing program jargon, these approaches are called "models" and their developers are called sponsors. Eventually, 22 sponsors were hired, and roughly 150 projects implemented their models, though the distribution of projects among the models was far from even.

These decisions, of course, were made over a dozen years ago, and the study that resulted from them is now completed. But the original concerns that gave rise to the study now face us again. How can the knowledge-production functions of the Follow Through program best be realized, and what role, if any, should the federal government play in coordinating the research? Today, it is NIE rather than OE who faces these questions. The NIE's ideas are still formative, and this paper as well as others presented here today are designed to help NIE develop those ideas. But the ideas appear to be sufficiently well-formed that at least three statements about them can be made. First, the NIE seems to feel that some federal coordination of at least some portion of the research would be useful. Second, the NIE seems to be interested in extending the thrust of the second version

of Follow Through by supporting the development and testing of new approaches to elementary education, independent of the Head Start link. Third, the NIE seems to be leaning toward a much smaller scale study than the second version's planned variation study--it will support smaller pilot tests of new approaches rather than multiple-site tests. These ideas, then, will constitute the third version of Follow Through to be discussed in this paper. To analyze these versions, I will now discuss them according to the seven study characteristics described above.

Programmatic Variations

Although the first version of Follow Through never existed, it is not too difficult to guess what it might have looked like. If one assumes, for example, as President Johnson apparently did, that Head Start wasn't doing as well as it could simply because there wasn't enough of it, then one might reasonably ask, how much more is needed? The Follow Through program would have provided a good vehicle for answering that question. Head Start graduates could be provided with Follow Through services in kindergarten only, kindergarten and first grade, kindergarten through second grade, and so on, thus providing a means for estimating the effects of different service durations. The study could also have included a "Head Start only" variation, so that the long-term benefits of Head Start itself could have been more carefully studied.

In addition to wondering about the benefits of different durations of service, our hypothetical program director in Washington might also have wondered how the Head Start and Follow Through programs could be coordinated locally. Follow Through would be offered in elementary schools, but Head Start is often offered at Head Start centers which are both physically and administratively separate from the schools. Yet if the Follow Through program is to really follow through on Head Start, one might desire certain programmatic consistencies--

materials, teaching styles, classroom groupings, or even staff continuity could be desired. At issue, then, would be how to achieve programmatic continuity in the face of administrative and geographic discontinuity. Thus, one might also want to compare different strategies for overcoming administrative and geographic barriers in order to create an elementary school program that would follow through on the Head Start program.

These would have been interesting and important aspects of educational programs to vary, but before considering them further, let's look at the programmatic aspects that were in fact varied in what I am calling the second version of Follow Through. In that version, the variations were called models, a term that is a little more appropriate than the term curricula, because the approaches themselves varied in a number of ways other than in their in-class activities and content. For example, they varied in their emphasis on classroom versus home activities and in the ways in which they integrated parents into their educational strategies. They are also models in the sense that many of them were more than prescriptions for what should happen; they were based on relatively comprehensive theories of child development and on philosophies of the purposes of education.

There have been several attempts to classify the 22 models eventually employed in Follow Through. Rather than reiterate them all here, I will describe instead the dimensions on which they vary. The most frequently mentioned dimension is that of how structured the classroom activities are. Those models called "structured" tend to consist of teacher-directed activities, and those labeled "open" tend to prefer that teachers respond to children's initiatives and capitalize on spontaneous events in providing their lessons. Although less frequently discussed as a dimension on which models vary, I should also point out that these models vary in how easily they can be taught to teachers. The open classroom approaches in general require teachers to do more thinking

on their feet and may require more creative teachers as well. Another dimension that is frequently mentioned is the relative emphasis on parent participation. All Follow Through projects are required to involve parents in the program, but the model sponsors vary in the extent to which they incorporate a role for parents in their models. A fourth dimension on which the models vary is their relative emphasis on different kinds of outcomes. Some are interested primarily in influencing the parents of children, others in influencing the children themselves; some are interested in children's specific skills, and others in their attitudes toward learning or their problem-solving skills. Finally, the models vary on a dimension that is rarely discussed but could be important, and that is their intensity. While some are only two-hour programs which allow teachers to do as they choose for the rest of the school day, others are full-day programs, so that children's entire educational experience is of a particular type.

With regard to the third version of Follow Through, the models NIE seems to be contemplating would incorporate a narrower range of variations than the second version. The emphasis is on approaches that require more than in-class modifications; that is, NIE wants approaches that require some restructuring of the organization and management of instruction in schools. But to the extent that the NIE-supported approaches do not vary in the role parents play and do not vary in profound theoretical or philosophical ways, the programmatic variations studied during the third version of Follow Through would not be as diverse as those studied during the second version.

Given these different versions of Follow Through, what can be said about the feasibility of altering education in the ways these versions are designed to study? The question may seem misplaced, since the studies could be designed to study implementation as well as outcomes, but it is an important question for three reasons. First, implementation is not easy to study; second, the experiences

of the second version of Follow Through as well as those of several other programs has been that implementation is even harder to accomplish than it is to measure. And those experiences lead to the third reason for asking this question: if implementation is not accomplished, then studies of outcomes have little value.

None of the models employed in the second version of Follow Through were easy to implement in the sense that the programmatic variations described in the first version might be, because the models require people to change their daily behavior. Teachers form habitual methods for asking questions, responding to questions, disciplining students, discussing students with parents, and so forth. In order for one of these models to be implemented, model sponsors spent enormous amounts of time working with teachers in their classrooms. The variations considered in my hypothetical first version have several advantages as far as implementation goes. It would be relatively easy to vary the duration of services, particularly if the nature of services was left to local discretion. And, although inter-agency cooperative agreements between Head Start centers and school district Follow Through projects might be difficult to reach, once they were reached, their implementation would follow accordingly. Neither of the programmatic variations discussed under the first version of Follow Through would require the continuous retraining of staff that was needed for the second version. Not only are they more feasible to implement but they are also cheaper, for program sponsors would not be needed.

Also to be considered with respect to feasibility is the probable transportability of the programmatic approaches that are studied. Presumably, if one of the variations tested were found to be more effective than others, there would be an interest in having other school districts adopt that approach. Once again, the variations studied under the hypothetical first version of Follow Through have an advantage. Not only are they easier to define, and hence easier to

communicate to others, but they do not require sponsors, and hence would be cheaper for others to adopt.

What about the third version of Follow Through? Many aspects of these new variations have not been sufficiently well defined that their feasibility could be determined. The NIE has suggested that the new approaches would not be sponsor-dependent, however, which may increase their transportability. On the other hand, someone has to develop these strategies, and someone has to implement them. If these two someones are the same--say, a particular school district, then their transportability would have to be assessed relative to the available research on district-to-district dissemination, and that research suggests that only the most clearly defined approaches are very transportable. Hence, the feasibility of testing the new approaches proposed by NIE may depend on the clarity of the approaches. To the extent that they lean toward philosophically-based approaches, such as those in the second version of Follow Through, they may not succeed; to the extent that they lean toward mechanical or logistical variations such as those described for the first version of Follow Through, perhaps they will.

Utility

The first and second versions of Follow Through are sufficiently different that one would expect them to serve quite different audiences. The first, for example, might be of as much interest to federal policy makers as to local ones. A study of the effects of different durations of service could help federal investors determine where their money would be spent best. Local administrations might also be able to use the data to decide where they should place their heaviest emphasis. Or, a local superintendent may use the data on alternative methods of coordinating elementary school services with Head Start services to design his own compensatory education program in his school district.

But there are also several things that could reduce the utility of the first version. First, since Head Start and Follow Through are administered by different local agencies, both might have to be involved in any decisions that might be based on the study. If one agency were interested, but the other were not, it may not be possible for the first agency to create the desired changes. Second, the two programs are not administered by the same federal agency; in fact, with the creation of a Department of Education, these programs now are not even administered by the same department. Separate federal administrative agencies could hinder local attempts to coordinate, by requiring audit trails on two distinct budgets and by producing regulations that are not complementary. And separate federal agencies could also mean that there is no clear federal audience for the research findings, or no administratively feasible way in which the federal agencies could act on the findings.

The utility of the second version of Follow Through did not suffer from such administrative problems, since the programmatic variations did not require more than one agency's involvement. But the fact that these variations proved to be so difficult to implement greatly decreased the utility of the study. That fact could not have been expected at the outset of the study, but today the experiences of Follow Through as well as a number of other demonstration and dissemination efforts suggest that probable ease of implementation should be a major consideration in estimating the utility of any proposed tests or comparisons of approaches, unless one assumes that the primary purpose of the study would be to inform the research community, so that more easily disseminated versions of the more effective models could eventually be developed. In fact, even though that use is hypothetically possible, and may yet prove to be a contribution this Follow Through study has made, it was never considered by the program administrators as a goal of the study. At different times during the course of the study, the

federal government expressed an interest in the benefits of Follow Through relative to no Follow Through without regard for the particular model; in model versus model, without regard for the type of sites they were implemented in; and in within-site Follow Through versus non-Follow Through, a set of comparisons which took into account both the type of model and the type of site.

The overall comparison of Follow Through versus no Follow Through could, in principle at least, have had value to federal administrators in determining whether the program as a whole had value and should be continued, but the fact of the models made such a comparison lack much meaning. On the other hand, the relative value of different models was not likely to be a topic of much interest to federal agencies, who continue to be chary of anything hinting of federal control over local curricula. Hence the design of this second version of Follow Through was not one that was likely to be useful to federal policy makers.

There was a time during the course of the study when the notion of a "consumer's guide to models" became popular. The idea was that the government would support the research, demonstrate the different kinds of effects that different models had, and thereby provide data that would be useful to local school districts. That concept proved unworkable for two reasons. The first has already been mentioned--the models would be too expensive for most school districts to implement without the help of a sizable federal grant. The second reason could only be known after the study was completed. The data indicated considerable outcome variation among projects implementing the same model. Few data were available on the quality of model implementation from site to site or on the ways in which local sites purposefully modified the models to adapt them to local needs; hence, the variations in outcomes could not be explained. In order for a study of programmatic variations to be useful to local program directors, some reasonable hypotheses about the relationship between particular site

characteristics and model outcomes would have to be available. Thus, the second version of Follow Through turned out to lack utility for either federal or local decision makers.

What about the third version of Follow Through? The NIE has suggested a desire that local superintendents be the primary audience, a proposal that seems reasonable, since they would be the people most likely to decide whether or not to adopt a new approach. They have also expressed a preference to avoid sponsors, thus making the approaches cheaper to transport to new settings. But there is one aspect to the utility question that has apparently not been considered: the extent to which anyone is actually looking for new approaches, or will be by the time the third version would be completed. We are in a period of decline. School enrollments are going down, and with them the tax bases that support schools. There is reason to believe that federal support for education will also decline over the next few years. For many school districts, the most pressing problems are not how to improve services, but rather how to maintain services in the face of reduced budgets. That cold fact may decrease the utility of a new version of Follow Through that entails the development and testing of new approaches.

Outcome Measurement

All three of these versions of Follow Through are designed to compare the relative benefits of some kind of programmatic variations on young disadvantaged children. And in education, the term "benefits" is almost synonymous with the phrase "test scores." Of course, the particular tests used don't have to fit the prevailing paradigm of the standardized norm-referenced test, but some form of evidence regarding children's academic standing would be needed for any of these versions. For example, even though the programmatic variations studied under the first version of Follow Through don't incorporate substantive or curricular

issues, they still are of interest because of their possible effects on children's achievement.

And in spite of the questions raised during the past few years about the quality and value of standardized achievement tests, these tests still present the most likely choice for measuring children's achievement in any of the versions of Follow Through. They might be augmented by a variety of other tests designed to measure attitudes or problem-solving abilities, but when the programmatic variations under consideration are administrative, without regard for curriculum, such as in the first version of Follow Through, the measures would necessarily have to be global indicators of achievement. The first version of Follow Through could also benefit from an investment in studies of the processes by which local Head Start and Follow Through projects coordinate their activities, but the findings from these studies would probably not be attended to unless there were sufficient program effects on test scores to warrant serious attention to the processes.

When the second version of Follow Through was being initiated, a great deal of thought went into the development and selection of outcome measures. A variety of contracts were let to university researchers and research firms to either evaluate existing instruments or to develop new ones. The primary evaluation contractor, Stanford Research Institute, also engaged in measurement development, and used a variety of strategies to involve model sponsors in their activities. Finally, when it was clear that these efforts were not bearing fruit, the administration convened a panel of Follow Through participants--researchers, sponsors, parents and project staff--to select the eventual test battery. The final instruments were all off-the-shelf norm-referenced tests, and were later supplemented by some classroom observation data on children's in-class behavior. Thus, in spite of extensive attention to the development and selection of outcome

measures, most participants in this version of Follow Through continued to feel that the measures did not adequately reflect the goals of many of the models under investigation.

There is no reason to believe that the outcome measures selected for a third version of Follow Through, regardless of the aspects of programs it was designed to study, would be substantially different--in either quality or variety--than it was 12 years ago when the second version of Follow Through was begun. And this state of affairs exists despite the fact that the weaknesses of various outcome measures were known at that time, and despite the fact that the intervening decade has witnessed a substantial investment of time and energy into measurement development.

Study Duration

The main reason I include study duration as an aspect to consider in trying to estimate the yield of a study is that things change over time. I already mentioned that these changes could reduce people's interest in the questions being addressed by the study. In addition, the passage of time permits a whole host of other events to occur that will confuse the interpretation of the data. Say the study takes six years, and takes place in 30 sites. During the 180 site-years of the study, old school buildings could burn or be condemned, causing the program in that school to relocate and causing major disruptions in the program's strategy. A major industry in town could close, thus drastically altering the town's population, and particularly its low income population. There could be a flood or a three-month-long teacher strike. There could be severe changes in the national economy, leading to either less funds allocated to projects, or to the deletion of some projects. Population shifts could cause some projects to become ineligible to participate, midway through

the study. Even a study that was designed with painful care at its outset would not be safe from the array of intervening events that can disrupt or substantially change the character of the programmatic variations that are being studied, so that when the study is finally completed, these events will have to be considered in the interpretation of the data.

There is yet another effect that the passage of time had on the second version of Follow Through, and that is that the outcomes were actually known long before the study was completed, for two reasons. First, even though the first two cohorts of children were not considered to have participated in fully implemented programs, they were tested at the end of third grade, while third cohort students were still in first and second grade. In addition, the third cohort of students was tested at the end of each grade level. All these data were analyzed and reported prior to the analysis of the third grade data on the third cohort. Thus, long before the final set of data were analyzed, those involved with the program knew that the results of their grand experiment were not exciting. The study was carried through to its conclusion in part simply to finish what had been started, in part in the vain hope that somehow the results would improve, and in part because there were no financial pressures to cut it off.

If the third version of Follow Through were based on the same premises as the second, it would require as long to complete as the second. That is, if the alternatives being compared are four-year programs (kindergarten through third grade), if a couple of years are given to projects at the outset to get their programs under way before they are evaluated, and if the final analysis isn't completed until a year after the final data are collected, then it will be seven years from the outset of the study to a final report. There is no reason to believe that the study will be immune from the several disruptive influences described above, and there is reason to believe there would be financial pressures

to cut it off before it was completed. 1981 does not seem to be a time to commence long-term studies in education.

Study Size

Studies such as those described for the first and second versions of Follow Through could be almost any size you wanted them to be. The problem is really one of deciding what size would be appropriate. If the first version, for example, were to compare alternative durations of service, it could be carried out by placing a Follow Through project in every community in the nation that has a Head Start center, and then assigning different communities to different durations of service; or it could be done within a single community, by assigning different children to receive different durations of service. If a middle-sized study were desired, it could be designed so that each community carried out a complete study, or so that each community implemented only one of the variations being studied.

If the first version were to compare alternative strategies for coordinating Head Start and Follow Through programs, its designers might have faced some of the problems faced by those who designed the second version of Follow Through. The study for the second version of Follow Through was actually designed after the programmatic variations were paired with projects. The design itself was a compromise between size and breadth. The designers wanted to incorporate as many of the models as possible into the study, yet several models were only being implemented by a few projects. Eventually they decided that, in order for a model to be studied, it should be implemented in at least five projects which offered kindergarten classes or at least three projects which did not. With these rules, the study sample included around a third of all the Follow Through projects, and around two thirds of all the models. While this strategy reduced evaluation costs considerably over what would have been necessary if the entire program were to be evaluated, it still was a very large and expensive study.

The third version of Follow Through will apparently be much smaller--indeed, it may be more similar to strategies employed by other demonstration programs than to the traditional Follow Through strategy. That is, if grants are given to individual projects to develop, test, and presumably disseminate later on, then the third version of Follow Through would be more similar to other federal demonstration programs than to the second version of Follow Through.

Inferential Problems

Most inferential problems come from the assignment process. That is, if children, schools, or projects are assigned to programmatic variations in any way other than randomly, it is difficult later on to determine whether observed outcome differences are due to the programmatic variations or due to the types of children, schools, or projects which happened to have been assigned to each program variation.

The first version of Follow Through would probably have such assignment problems, since many individual sites may only be capable of implementing certain variations. For example, some school districts don't routinely offer kindergarten classes, some Head Start centers don't routinely offer Head Start to both three- and four-year-olds. Some Head Start centers are physically housed in school district buildings, and even administered by school districts, while others are separate. A study along the lines of my first version of Follow Through could take one of two routes. The first route would be to assign sites to conditions they can implement, and try to live with the inferential problems that will result. The other would be to design several smaller-scale studies in which sites are first grouped according to their existing organizational arrangements, and then whatever variations are feasible within each arrangement are compared within each group of sites. The second choice is not necessarily a solution to the assignment problem, since some categories of sites may be too small to conduct comparative

studies in, and other categories may be constituted such that there are few variations worth comparing in them.

One could argue that the effects of naturally occurring variations in site organizational characteristics could be studied in itself, and could yield some important information. But if these natural variations are not amenable to change, then it's not clear how anyone could use the findings of such a study to improve the education of young disadvantaged children.

But problems of assignment are not the only source of inferential problems in the first version of Follow Through. The entire strategy embodies two assumptions which I have not mentioned yet, but which could be significant when it comes time to draw inferences from the study. The two assumptions are related. One is that the original Head Start program, on which the Follow Through program is to build, is as good as it can be; the second is that it will remain unchanged for the duration of the study. Suppose neither of these is true. Suppose instead that Head Start is not as good as it could be, that its administrators know that, and that its administrators continue to conscientiously modify and improve the program over time. Meantime, its earlier graduates are participating in the study of combined effects of Head Start and Follow Through, which will not be completed until several years after the children leave the Head Start program. By the time they complete the Follow Through program and the data are analyzed, the Head Start program may have changed considerably, thus raising questions about whether the study's results would have occurred if the children had participated in the current version of Head Start. That question casts doubts on the value of the entire study and on the kinds of inferences that can be justifiably drawn from the study.

Since the second version of Follow Through did in fact occur, the inferential problems raised by the assignment strategies as well as a number of other

inferential problems have received considerable attention in the literature. I have already mentioned one of the problems--the variation in outcome among sites implementing the same model. That problem, of course, could not have been anticipated at the outset of the study, but it is a sufficiently pervasive problem that it would have to be considered in the development of any new planned variation studies. Two other inferential problems have also received attention. One of these is the fact that the outcome measures were not designed to reflect the outcomes of interest to several of the models. The measures have been criticized both for their content and for their format, and several authors have suggested that these inadequacies make it impossible to assess the real effects of the models. The other problem that has received attention is the fact that models were not assigned to projects; rather, projects were allowed to select the model they wanted to implement. This led to a very awkward pattern of models, at least from a research perspective. One model had over 20 projects, others only one. One model had around 10 projects, but eight of them were in the Northeast. There are many who argue that the potential for biases in this method of linking projects with models makes it impossible to tell whether the observed effects are due to the models themselves or due to some unknown characteristics of projects which predisposed them to select these types of models in the first place. Of course, one could argue equally strongly that imposing models on projects that didn't want them would create even greater inferential problems, and further, since the findings are intended to generalize to other projects who might voluntarily adopt the models in the future, the outcomes <u>should</u> reflect the effects of models in sites that voluntarily adopt them.

There is another aspect of the assignment issue that has received attention as well, and that is the way in which comparison groups of children were selected

within each project. Comparison groups were selected after Follow Through children were, and they were selected to be as similar as possible to the Follow Through children. But the degree to which they were similar varied from site to site, and these variations could have contributed to the observed site variations in outcomes. Although statistical adjustments were used to account for initial differences between the groups, and the comparisons were carried out separately for children with and without preschool, one never can be sure that analytic strategies are sufficient to counteract initial differences between groups.

Finally, there is one inferential problem that has not been discussed much but needs to be mentioned here, and that is the fact that few of the models exist independent of their sponsors, and that sponsors differ in their ability to generate enthusiasm about their models, to explain clearly what their models entail, and to hire and train staff members who will in turn train teachers. They differ not only in individual personalities, but also in their institutional characteristics. These characteristics make it impossible to estimate what effects the models might have if others were assisting projects with them--or even to tell whether anyone else could assist projects to implement the same models.

The third version of Follow Through would avoid many of the inferential problems inherent in one or the other of the first two versions. For example, since it does not require a linkage to Head Start, it avoids problems of inference that could arise if the Head Start program were to change during the course of the study. And since it would not rely on sponsors, it would avoid the problems of interpreting whether observed effects were due to models or to their sponsors. The most serious inferential problem it would probably face, and one that cannot be avoided, is the problem raised by variations in outcomes among projects implementing a common model. The version currently of interest to the NIE would

not include replications of approaches, as I understand it right now, but that hardly solves the problem--indeed it exacerbates it. Whether one explains the site variation in the second version of Follow Through as due to measurement error, to implementation variations, to variations in the equivalence of Follow Through and non-Follow Through children, or simply to variations in inherent characteristics of the sites, these competing explanations for outcomes will be present in the third version of Follow Through, regardless of the number of replications of each approach that are tested.

Cost-Benefit Implications

It is not a simple matter to estimate the benefits of a study relative to its costs. Costs can be at least broadly estimated by the size of the study, but benefits can vary as a function of the several things discussed above as well as several others, including serendipity. If the second version of Follow Through were to be repeated and its size held constant, the benefits from the study could be increased by certain modifications in the study design. For example, the number of models could be reduced to three or four and the replications of each model increased to 15 or 20. The greater number of replications, coupled with studies of implementation might make it possible to determine the nature of interactions between projects and model, the extent to which a model may have an average site effect that is recognizable over the site variations, and so forth.

A number of arguments can be made for small scale studies other than arguments of cost. Smaller studies allow for more control over events that may render data uninterpretable. For example, the processes of assignment can be more controlled, funds can be retained for studies of program processes, children who leave the program can be followed and tested, so that the effects of attrition can be better estimated.

Yet while all of these arguments are valid, small scale studies introduce their own set of inferential problems too. They may only be possible in relatively sophisticated communities, with universities nearby, so that the findings may not generalize to other places. Or the fact of having more than one program variation within a community could heighten people's awareness of the experimental aspect of their services and create effects that would not have occurred if the site were implementing only one of the variations.

The relationship between costs and benefits may also depend on the audience. If, for example, the audience is federal administrators or law makers, who would be most likely to use the findings to determine the future funding structure and regulations for the program, perhaps smaller, carefully controlled studies of the regulatable aspects of the program would be sufficient. If, on the other hand, the audience is to be local administrators, the study may need to be considerably larger, so that a broader array of site characteristics can be taken into account. And if the larger study were to be undertaken, its benefit to this audience may depend heavily on the extent to which the relationship between site characteristics and programmatic outcomes can be explicated during the course of the study.

These comments on the costs and benefits of Follow Through, however, do not take into account another kind of benefit the study has provided. In the decade since the study began, researchers and evaluators have become considerably more sophisticated with regard to conducting studies outside the laboratory. Our awareness of the problems that can arise from non-random assignment has increased, our statistical strategies for adjusting scores have improved, our familiarity with the range of competing hypotheses that can account for differences among groups has increased, and so on. Much of this increased sophistication would have occurred without the presence of Follow Through, but a great deal of it came

about because of Follow Through. The size and duration of the study allowed a number of different evaluators to participate in it, and many of them have shared their knowledge and experiences with their colleagues. The study has proven to be a valuable "case-in-point" for articles and textbooks dealing with topics from disadvantagment to educational psychology to research methodology. Unfortunately, it is so often used as a negative example in discussions of research methodology that it has come to be viewed as worse than it actually was. But these discussions are valuable to the profession, and constitute a legitimate benefit that the study has had. In addition, by virtue of the sponsorship strategy, we have learned a great deal about what is involved in converting theoretical principles and laboratory findings into classroom practices and that knowledge is something anyone who is concerned about the applicability of educational research needs to have.

Conclusions and Recommendations

This year, the evaluation funds traditionally used for Follow Through's planned variation study are being transferred to the National Institute of Education. NIE now has both the opportunity and the responsibility to develop a strategy for spending that money that will result in a reasonable knowledge yield, given the funding level. NIE seems to be moving in the direction of testing new educational strategies in a way that differs slightly from the other two versions of Follow Through that I have discussed above. Although the details of NIE's plan have not been worked out, they are sufficiently clear that some statements regarding their probable yield could be made. My concluding statement about the third version is a rather long one, so I will state the punch line now: I don't believe the third version will have any more yield than the second did, and in fact will probably have considerably less. Having said that, I will now provide eight reasons why, and then propose an alternative

third version.

First, it is not likely that there are any other approaches that would be radically different from the 22 that were initiated for the original Follow Through study. In fact, in 1975 the NIE attempted to find new approaches to compensatory education, and funded four groups to develop their approaches. One approach suggested moving compensatory education from early elementary school to secondary school, an approach that could not be tested in Follow Through. One proposed working more with parents and the community, an approach that has already been tested in Follow Through. One proposed radical architectural changes that would be too costly to implement for purposes of an experiemnt. And the fourth approach proposed cross-age tutoring, a strategy that could be tested in Follow Through and probably has already been tried by some of the model sponsors. If we assume Follow Through will continue to serve young disadvantaged children within existing educational agencies and buildings, with about one teacher for every 30 children, then there simply aren't many ways it can be done, and most of those ways already exist in Follow Through.

Second, the variations studied in the extant version of Follow Through were more various and more intense than any other educational variations ever studied. They were four-year programs, hence longer than most other programs studied; they served children by modifying their regular classroom experience, rather than pulling them out of class for an hour a day, and hence were more complete than most other interventions studied; and they were more different from one another than most other programs are that are compared. Finally, the teachers had the benefit of continual sponsor training and support for the duration of the study, whereas many programs are considered to exist after a week of training at the beginning of the year. When programs of this diversity, this duration, this comprehensiveness, and this intensity of staff support fail to show anything more

than trivial effects, one cannot attribute those effects solely to design problems. The originators of the Follow Through study expected the data to display "inter-ocular effects"--effects that hit you right between the eyes. They had every reason to expect the outcomes to be dramatic. Today, we know better. It would be very hard to justify an argument that a set of less diverse and less intense alternatives would demonstrate more dramatic or more consistent effects.

Third, if ever there is or was a time when school districts were not interested in trying new innovative approaches to education, it will be the decade of the eighties. We have entered a period of severe decline. Enrollments are declining and budgets are declining. Several state legislatures have imposed tax caps, and others probably will. The prevailing wisdom is that federal support for education will also diminish. What will concern most school districts, then, will not be how to improve education, but rather how to maintain it at reduced costs. That means the utility of research on the relative benefits of innovations will not be great.

Fourth, in spite of the tremendous attention given to evaluating and refining outcome measures and to developing new ones, there are still very few acceptable instruments available. As long as these instruments can still be criticized for their lack of program sensitivity and their possible racial and ethnic biases, any evaluations of program outcomes are suspect, and evaluations of programs serving disadvantaged children are even more suspect.

Fifth, the study would, like the other two versions of Follow Through that I have described, take too long. With a new cost-conscious administration in Washington, it would be foolhardy to begin a study which may not be possible to complete because budgets are cut, or worse yet, because the entire program is discontinued.

Sixth, new planned variation studies would pose the same problems of site

that the first two versions I have discussed posed. The studies would have to be massive in order to stratify on all relevant site characteristics and to include sufficient replications, or they would have to be self-contained within individual projects, thus requiring the presence of sophisticated research talent on-site, and thus raising questions of generalizability of the findings.

Seventh, if studies involve individual projects as both developers and testers of new strategies, with no replications, they will present the already well-documented problems that other demonstration programs have, that these projects don't replicate elsewhere any better than sponsored models do.

Eighth, any new study that tests approaches is highly likely to have many of the same inferential problems that the second version of Follow Through had. The groups of children whose eventual achievement will be compared will probably be dissimilar in other respects, the measures will not be adequate to tap program goals, attrition will reach 80% by the end of the study, there will be a variety of intervening influences to have to consider when interpreting the results, and the likelihood of replicating the program or its effects elsewhere will be unknown.

The eight arguments listed above all suggest that an NIE-sponsored third version of Follow Through would have limited benefits, regardless of its costs. I have posed these arguments in their strongest and bluntest form, and perhaps unnecessarily so. But I did so for a reason, and that reason is to shake the notion of testing new approaches out of the heads in which it seems to be firmly planted. Let me quickly add that I strongly endorse the concept of federally-coordinated research in a demonstration program, and think that that approach will eventually produce much greater benefits than the research and development activities typically supported by demonstration programs. On the assumption that I have made the case that the notion of developing and testing educational approaches so that they can be adopted or adapted elsewhere is simply not feasible--that these studies are too complicated, too costly, and too uninterpretable to have

value--I will turn my attention now to another way in which the third version of Follow Through could be done.

My alternative for the third version of Follow Through is much simpler to comprehend and to carry out, and I believe it would yield even more useful research findings. Under this version, the program continues to exist in roughly the same form as it has in the past. Projects are, for the most part, implementing models with the aid of model sponsors. The particular project-sponsor affiliations may change, some projects or sponsors may drop out, and some new ones might enter. Sponsors also continue to study implementation, revise their models, improve their training strategies, and develop new outcome measures. All of these activities continue to be supported by the Follow Through program.

NIE, on the other hand, would support research projects that would increase our understanding of all the areas which Follow Through comprises: disadvantagement, the process of innovation, the process of implementation, the social settings in which innovations meet disadvantaged children, the effects of federal regulations, and the nature of incentives for change in school systems. Let me list some examples to demonstrate what some of these studies might be about.

1. Studies of the nature and implications of disadvantage.

President Johnson's original war on poverty was based in large part on the premise that education should be a major battlefield in the war. It still is not clear that education should have that status, nor more particularly that early education should. There is a whole side to Follow Through that was ignored during the conduct of the second version of Follow Through, and it is the side most Follow Through program staff would argue is the most important side. That side of Follow Through is not concerned about planned variation; it is concerned about offering comprehensive services to poor children, making their parents more self-sufficient, and establishing stronger ties between schools and the poor communities they serve.

These goals are far-reaching and dramatic, but whether they have been met has never been determined. Indeed, we have no evidence regarding whether schools are even reasonable places in which to try to reach these goals. Intensive case studies of individual families in individual Follow Through communities could offer a variety of valuable insights into the nature of disadvantagement, how schools are perceived by poor families, and how Follow Through in particular interacts with and influences poor families.

2. The organization and delivery of non-instructional services. I could also imagine a number of studies designed to serve local administrators--not studies of what programs work, but studies of how to improve the administration of services. We know little about the nature of inter-agency agreements that are necessary to coordinate services, or the best ways to reduce service delivery costs without seriously impairing the quality of those services. Nor do we know how schools obtain non-instructional services, what they do when no such services are available in the community, or what they are capable of doing under different circumstances.

3. Federal regulations. Several aspects of Follow Through are found in a number of other federal programs as well. Requirements such as the "supplement--don't supplant" rule have become part of local school vocabulary, and parent involvement is almost universally required. Yet different programs enforce these rules in slightly different ways, and permit more or less site autonomy in how the provisions are met. The result is that there is a considerable amount of naturally-occurring variation in the extent to which these rules are met, bent, or completely ignored. Understanding the organizational causes for these variations, and the circumstances under which such rules enhance program quality or hinder it could be useful not only to Follow Through administrators but also to administrators of the other federal programs. These administrators are under constant pressure to reduce

regulatory requirements, and may have little knowledge of what effects their requirements actually have on school districts.

4. Implementation. The literature on implementation has grown tremendously since the second version of Follow Through began, but Follow Through itself has contributed little, relative to what it could. What do sponsors actually do? How do they operate? And how do they compare with other change agent strategies? If the program begins to fund new projects in the next few years, it might be possible to study the processes by which projects and sponsors originally come together. I, for example, would be interested in knowing whether either the rationale used by the project when selecting a model or the points emphasized by sponsors when selling their models turn out to be valid once the model is actually implemented in a given school district.

5. How the models work. It would be extremely valuable to conduct ethnographic studies of classrooms implementing different models, to learn more about what these models are really like. The classroom observations conducted during the original study were helpful in delineating model differences in such areas as children's question-asking and teacher question-asking, but still could not give much flavor of the real dynamics of different kinds of classrooms. Since many projects will probably continue in the program using the same models, these studies could be carried out now.

These suggestions are offered merely by way of emphasizing that there are a great many important educational questions that can be answered that do not involve the testing of educational models, and they are questions that can best be addressed through a federally-coordinated research program. After three to five years of research on the nature of models, the processes of implementation, and the characteristics of other aspects of the program, we may be in a better position to begin a controlled study of program effects--or at least to know what things

we would want to control if we were to test the effectiveness of a particular educational approach.