DOCUMENT RESUME

ED 254 337 PS 014 958

AUTHOR Gross, Beatrice; Gross, Ronald
TITLE Frontiers of Research and Evaluation in Compensatory Education. A Report of the Follow Through Planning Conference "Documentation of School Improvement Efforts: Some Technical Issues and Future Research Agenda" (Pittsburgh, Pennsylvania, March 12-13, 1981).
SPONS AGENC National Inst. of Education (ED), Washington, DC.
PUB DATE 82
NOTE 52p.; For individual conference papers, see ED 221 557, ED 226 453, ED 242 427, ED 243 585-587, ED 244 723, ED 244 738, ED 245 791, and ED 245 795.
PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Compensatory Education; *Early Childhood Education; *Economically Disadvantaged; *Educational Facilities Improvement; Evaluation Methods; Evaluators; *Program Evaluation; Research Design; Research Needs; *Research Problems; Research Proposals; Young Children
IDENTIFIERS *Project Follow Through

ABSTRACT

This conference report identifies research needs in evaluating and documenting large scale school improvement efforts to serve disadvantaged populations. Summaries of the conference presentations are provided, grouped into three sections. The first section examines the basis of conducting evaluations. Several basic assumptions are challenged, among them that well-planned innovative programs have an appreciable effect, that research findings influence educational decisions, and that the use of scientific inquiry is a valid basis for social change. It is suggested that a "judicial method" can involve school staff in constant controversy and keep them learning as they weigh "trial" evidence. The second section discusses methods of knowing whether programs are being implemented. Discussions of a method to calibrate degrees of implementation by teachers, the "banking" of a retrievable group of generalized outcome measures (which would serve Follow Through programs that are ill-served by standardized achievement tests), and the diverse requirements for federal administrators are also included. Uses and misuses of tests and the inadequacies of the testing system for language-minority students are discussed in the third section. The report concludes with discussion, conclusions, and recommendations by conference members; a list of participants; and a list of papers presented. (AS)

FRONTIERS OF RESEARCH AND EVALUATION IN COMPENSATORY EDUCATION

A report of the Follow Through Planning Conference
"Documentation of School Improvement Efforts:
Some Technical Issues and Future Research Agenda"

by

Beatrice Gross and Ronald Gross

National Institute of Education
1982

TABLE OF CONTENTS

This report was prepared with funding from the National Institute of Education, U. S. Department of Education. The opinions expressed do not necessarily reflect the positions or policies of NIE or ED.

INTRODUCTION:

Summoning the "Collective Wisdom of Top Thinkers..."

Using this phrase to characterize the conference reported in the following pages, Professor Margaret Wang, convener of the conference, had chosen her words with care. And they were justified.

The conference brought together an outstanding array of educational researchers, practitioners and policy-makers. Their challenge, posed by the National Institute of Education, was to identify research needs in evaluating and documenting large-scale school improvement efforts to serve disadvantaged populations. In short: how could the best scientific and humanistic know-how available today assure that further public investments in compensatory education would bring solid results and insights.

The conference was convened at the Learning Research and Development Center (LRDC) of the University of Pittsburgh by Professor Wang, Sponsor-Director of LRDC's Follow Through Model, on March 12-13, 1981. Twenty-five experts, representing the frontiers of evaluation research in the country today, met to discuss issues in the design of supporting research for the planning and development of future Follow Through Projects.

The official conference title was "Documentation of School Improvement Efforts: Some Technical Issues and Future Research Agenda." But the real "working" title might well have been "This Time, Let's Do It Better."

Co-sponsored by the National Institute of Education (NIE) and the LRDC as part of the "second strand" of NIE's Follow Through Planning Conferences, conference activity started back in 1980 when the Office of Elementary and Secondary Education and the Office of Educational Research and Improvement authorized the NIE to embark on long-term research and demonstration projects to try out alternative approaches to educating disadvantaged primary school students. NIE used 1981 to plan its activities. Advice, recommendations and input were sought from a wide range of individuals. The Institute commissioned 44 papers and arranged invitational conferences in Portland, Oregon and Austin, Texas to address evaluation concerns and in Pittsburgh, Pennsylvania (the report of which follows) to address research and evaluation.[1] Subsequently, on June 10, 1981, a Request For Proposal (RFP) was issued by the NIE and, after a nationwide competition, four contracts were awarded on September 30, 1981 to Oakland, California, Napa, California, Detroit, Michigan and Cotopaxi, Colorado.

The experts invited to the Pittsburgh conference were all told that:

> Since NIE is planning to embark on a 15-20 year program, recommendations are needed for both short-term and long-term activities. The LRDC and NIE conference will mostly address long-term activities and methods for continually receiving recommendations.

Prior to the meeting, participants were sent all conference papers and abstracts, and a copy of the NIE Planning Paper of October 1980. In their invitational letter dated March 2, 1981, they were updated on NIE's current thinking as follows:

> As a result of the planning conferences in Portland and Austin, NIE has refined its tentative plan for its involvement with pilot Follow-Through projects. Basically, we are planning to seek low-cost school-wide approaches toward educating disadvantaged children. This implies an emphasis on methods for the management of instruction rather than curriculum development. The conceived evaluation activities would document the implementation of each approach over a three to five-year period and would be designed to serve potential adoption/adaption sites in their efforts to determine whether to implement a similar program at their site.

[1] A synthesis of these three conferences was published by the NIE. "Planning For Follow Through Research and Development" includes a short history of Follow Through and is available by writing to Charles Stalford, National Institute of Education, Stop 9, 1200 19th Street, N.W., Washington, D.C. 20208-1101.

> From the LRCD-NIE Conference, NIE is seeking sets of written recommendations on:
>
> 1. the evaluation of NIE-funded pilot Follow-Through projects
> 2. needed evaluation methodology research
> 3. needed instrumentation research and
> 4. compensatory education research.
>
> Recommendations on conducting evaluations are needed for projects to commence during the next school year. Research recommendations are needed to provide support for future program and evaluation activities.

They Came With Baggage

The participants shared another common background over and above the challenge presented them by the NIE. Many of them had been involved with Follow Through from its inception, and the others had been well acquainted with its mission and its struggles. The relevance of this background to their responses to the NIE challenge was well expressed by Prof. Wang:

> Those of us who have been affiliated with Follow Through since its early years probably can recall the heated debates we had for years over the one central question: How do we go about identifying the best models of early childhood education? This dispute prevailed during Follow Through's first decade and involved a range of topics which included the specification of the goals and objectives of the National Follow Through Program, the expected outcomes of the Program, the best measures to assess those outcomes, and the ill-fated match between the measures used in the national FT evaluation and the goals and objectives of specific model programs. The design problems we discussed were concerned mostly with the non-comparability of the control groups used in the national evaluation study of FT. We were very much preoccupied with the classic question: If you get good results, how can you be sure it's because of your program if you don't have adequate comparison groups?
>
> Most of [the discussions] were based on the "givens" on how to evaluate Follow Through at the time. The accepted design was based on the treatment-outcome paradigm, and the sole purpose was to compare the relative effects of the various model programs on outcome measures. For several years, we tried to utilize this design to solve our evaluation problems. As it turned out, the solutions

> never came, mostly because we were asking the wrong questions..."

Describing how evaluators were sidetracked by taking the technological route, she continued,

> those of us who were worrying about the technological and methodological issues did not give much thought to such matters as the conditions under which we would have to operationalize our evaluation design. We were totally immersed in the challenge of coming up with methodologies and designs that could answer Follow Through's basic evaluation questions, and we were totally ignorant of the facts of life of implementing and studying school change. We tried to solve our problems through the use of sophisticated statistical methods of data analysis which were technologically elegant but which made our lives even more difficult as we attempted to tease out the relevant information needed to discover which model programs were more effective.

ADVANCING THE FRONTIERS

Clearly, from the syntheses of the presentations that follow, the field of evaluation is getting older and wiser. What was once thought to be a simple task -- that of deciding which of several programs is better, why, and how another can be made in its image -- is now seen as highly problematical. It was evident listening to the researchers who spoke at this meeting that they were among those in the field who had shifted from what Wang called the "fidelity perspective" to the "adaptive perspective." The problems of evaluation designs described in part by Wang were probed and analyzed at this meeting by "Individuals not only recognized for their talent and contributions to advanced research, but also for their leadership in challenging colleagues to make evaluation research useful and relevant to school improvement," in the words of convener Wang. These presenters were:

Gene V. Glass, University of Colorado at Boulder.

Leigh Burstein, University of California at Los Angeles.

Garry McDaniels, U.S. General Accounting Office, Washington, D.C.

Ernest House, University of Illinois, Urbana.

Thomas McNamara, Philadelphia School District.

Susan Loucks, The Network Inc., Andover, Mass.

Chad Ellett, University of Georgia at Athens.

Walter Haney, The Huron Institute (Cambridge, MA).

J. Ward Keesling, System Development Corporation (Santa Monica, CA).

Dalton Miller-Jones, University of Massachusetts at Amherst.

Ernest Bernal, Creative Education Enterprises (Austin, TX).

Starting off the first day's provocative presentations, Gene Glass challenged several basic assumptions of the field: that well-planned innovative programs have an appreciable effect; that research findings influence educational decisions; and that we should be looking for one or two "right programs" for all children and get school people to use them.

But even these radical challenges to the conventional wisdom were exceeded by Ernest House's disputing the use of scientific inquiry itself as the basis for social change.

Having heard such sharp affronts to some basic axioms, the conferees took in stride both Leigh Burstein's contention that because children are in a dynamic environment, any conclusions reached about one group cannot reliably predict what will happen the next time; and Thomas McNamara's similar perception about the effects of the dynamic school climate on staff. McNamara recommended the "judicial method" which would involve staff in a constant controversy and keep them learning as they weighed the "trial" evidence. This presaged Susan Loucks' description of the levels of increasing sophistication leading to full implementation of a program in the classroom.

Chad D. Ellett contended that at present we can't claim to know anything about the results of programs based on their evaluations, because we still do not have the tools to know if the programs are really being implemented. But he did suggest some directions that we might take to devise these tools.

Also concerned about the lack of instrumentation was J. Ward Keesling but he focused on "banking" a retrievable group of generalizable outcome measures which would serve the hitherto unserved Follow-Through programs that were ill-served by standardized achievement tests.

Garry L. McDaniels explained why reports to Congress required information that was not always useful to people working in the field, why no one investigator should be expected to handle.the diverse requirements needed for federal administrators.

Finally, the nuts and bolts uses and misuses of tests were discussed by Ernest Bernal, who pointed out the inadequacies of the testing system for language-minority students; by Walter Haney, who would like to see tests teach, and be used by teachers who want to zero in on their children's needs; and by Dalton Miller-Jones who exposed the logical inconsistencies in the present standardized tests, and urged tests that will help us understand the cognitive processes of children so they can be taught successfully rather than merely being sorted into winners and losers.

Examining the Basis of Our Judgements

USEFUL EVALUATIONS

> "The art of teaching must not be subordinated to the technology of mass testing."
>
> Gene Glass

"If it were not that so many people are intimidated by the evaluators' methods, the arbitrary authority of evaluators would more quickly be seen as illegitimate. The truth is, we evaluators don't know much and we don't know how to use what we do know," said Gene Glass, from the University of Colorado at Boulder. Long regarded as an expert in the uses of research and research analysis, Glass has been working during the past few years on a project "summarizing and integrating research findings of different educational treatments."

He has concluded that teachers decide matters of curriculum and approach on the basis of complicated understandings, beliefs, motives and wishes. Research findings have little influence on these decisions, according to Glass, and that's just as well, since in his experience, even well-planned innovative programs don't have an appreciable effect.

Glass drove home his point with data about a wide range of interventions: psychotherapy, programmed instruction, drug treatment for hyper-activity, treatment of learning disabilities, tutoring programs, mainstreaming, Transcendental Meditation, behavioral treatment of structuring, and perceptual-motor training. "What I have found," said Glass, "is that in the majority of cases, the variability in any experiment is, on the average, twice as great as the improvements the experiments show. Also, the mean is only half as great as the variability, and the odds run about three to ten that you will find the cheaper control treatment showing better than the innovation being tested. In essence, then, experiments don't have large effects. We can't reliably predict that A will be better than B, or B better than A." In fact, Glass finds the odds about even that a new treatment might be <u>worse</u> than what was going on before.

In education, therefore, even if decisions to adopt innovative programs *were* actually made on the basis of such things as matrix sampling, logistic item models, factor analysis and the like, teachers whould not have had a good reason to opt for the programs.

"Since the conditions of the Follow Through programs that were evaluated were frequently not known or were not consistent across programs of the same model, anyone who aspired to replicate the 'successful' programs was bound to be disappointed," said Glass.

"It would not have been rational for schools or teachers to adapt even the seemingly successful Follow Through programs on the basis of existing statistical data. Even if one overlooked the fact that we knew very little about the circumstances of the model projects, the results of what *was* tested (which of course did not attempt to translate the complex subtle notions of child development and goals of education into mass tests) did not support the innovations sufficiently."

The problem isn't one which can be solved by investing large amounts of money in synthesizing test results.

"Teachers don't need and don't use statistical findings of experiments when deciding how best to educate children. They do want to know whether the method is consistent with their views of themselves as professionals, whether the program treats pupils as though they were robots, delicate flowers, or children of God."

To do better in the future, Glass advocates "evaluations that emphasize description (principally *qualitative*) for informed choice. Models should be described in terms that people consider personally significant when they choose a particular profession for themselves or a school for their children. Technocratic, behavioristic and anti-democratic language should be avoided. An ethnographic or case-study approach to evaluation should be adopted in place of a quantitative, experimental field trial. What teachers need to make informed decisions are:

- Some coherent, detailed portrayals of life in school for pupils, teachers and parents as it is colored and shaped by allegiance to a particular Follow Through model,

- Some portrayals by disinterested, expert ethnographers with at least two years on-site for data collection and,

- Some portrayals focused on a broad range of concerns including the model's philosophy, its history (since its future must be projected), techniques, financial and psychic-costs, side-effects and after-effects, the roles it requires people to play, its potential for a favorable evolution, and the like."

Concluded Glass: "Our evaluations should not aspire to discover the one or two right programs for all children and get everyone to follow the prescription. We need evaluations that will lead to adoptions by school people, who can make an informed choice, based on their goals and philosophies and the nature of their districts."

SCIENTIFIC AND HUMANISTIC EVALUATIONS

> "For the guidance of future human action, one would choose the humanistic study over the scientific one."
>
> Ernest House

Two children from the same family go to the same school and are exposed to the same programs (Distar for reading, IPI for math). One is stimulated, the other bored. How can that be? Becuse, explained researcher Ernest House, from CIRCE, the University of Illinois and the father of the children, they had different personalities and different learning styles. Because too, the vivacious teacher who had iniciated the program and had taught his older child had left the program and the new teacher who taught his younger child was not as lively. Also because the program itself had lost its glowing promise by the time his second child was enrolled.

These reasons cited by House for why his two children reacted as they did explain in small part why no program, regardless of how specifically delineated, is the same for every child. Even more variations occur when programs are implemented in different schools, in different towns with different socio-economic groups, by different teachers, etc. Given these variables we must question what we accomplish when we collect and quantify data across classes, schools, communities. What are the implications of this insight for testing and evaluation? According to House, they suggest we must critically evaluate why we do what we do and whether we should change our thrust in research.

According to House, social scientists are on the whole, solidly in the tradition of Leonardo, Copernicus, Galileo and Newton -- that is, in the "scientific tradition." They are looking for a mathematically measurable answer to formulate a universal law of reality.

Modern science has three assumptions:

- that every question has one and only one true answer and if one doesn't arrive at the one true answer, one has asked the wrong question, for the right one will yield the right answer

- that there is one method for discovering the answer and the method is rational in character

- and that the answers discovered by such a method are true universally for all people in all times and that truth is not relative in any way.

House disputes the idea that physical and social reality are similar. He believes that the "scientifically" designed findings of the Follow Through evaluations done to date are inadequate "even though elaborate quantitive methods were employed" not simply because the wrong methods were employed but because the basic assumption is that scientific inquiry can be employed for this purpose, is incorrect. House would substitute for scientific inquiry a report of the experiences of those involved, even though those experiences may be "biased, subjective and undisciplined."

Citing the writings of Vico, a Renaissance thinker who proposed that there was no point in behaving as if human nature is unchanging, House offered a disciplined alternative paradigm to the scientific one provided by Galileo. In this alternative view, individuals and their actions are seen in terms of their intentions and purposes.

House recommended the one Follow Through evaluation which was not based on the "scientific method" but was, rather, an excellent example of the humanistic model of inquiry. The Bank Street study, written by Zimiles and Mayer (1980) is subjective and can be accused of being biased, but, in House's opinion it gives far more valuable information than the spare scientific studies most frequently cited in the literature. So, for all its weaknesses, the length being a big one, House recommended that for some purposes and in some situations, the humanistic study may be preferable to the scientific study.

INVESTIGATING SOCIAL PFOGRAMS WHEN INDIVIDUALS BELONG TO A VARIETY OF GROUPS OVER TIME

> "Researchers must remember that they are dealing with dynamic subjects in a changing environment when they set about their work, because these dynamics limit the programs' predictability."
>
> Leigh Burstein

How much does the context (the nature of the kids, their out-of-school experiences, the abilities and personality of their peer group, etc.) affect the validity of the evaluation? According to Leigh Burstein of the University of California, Los Angeles, researchers must be mindful that students and schools are in constant transition, and these dynamic properties will affect research results. Therefore, we must know:

- If the program was actually implemented.
- The adjustment teachers and students had to make.
- The effect of the reform on the social system of the school itself (did it work at cross purposes or blend smoothly).
- The circumstances and nature of the children involved in the programs.
- The effects of the composition, size, ability and personality of the group.
- The different effects different programs had on different kids, how long they lasted, and if children outside the program behaved differently.
- What kinds of relationships the programs engendered -- i.e., cooperative, competitive.

- The educational achievement (both short and long term .

- The attitudes engendered towards self and schooling; initiative, independence, adaptability, etc.,the well-being that apparently resulted.

- The possible effects of the shift from one learning environment to another (what Burstein calls discontinuity).

- Any changes in school attendance, special education placement, grade retentions, etc., that resulted.

- Did the students who participated find it difficult to adapt to a new instructional style afterwards -- so that the discontinuity of experience was seen in the long run to be detrimental even though the immediate result was that the program fostered better skills as a result of improved instruction. (In such a situation, districts might have to consider keeping the same system and teacher throughout at least the first three years of schooling as they do in Sweden.)

Wrapping up his argument, Burstein asserted that since "program elements are inherently interrelated and their interface, linkages, and dependencies are at the heart of a sound understanding of school reform efforts," "better conceptualization, design, instrumentation and analyses will improve the process only marginally unless refinements are directed towards understanding both program elements and their interrelationships by combining the focus on educational and social processes with multiple investigations from diverse perspectives."

PUTTING AN INNOVATION "ON TRIAL"

> "The judicial approach is . . . particularly relevant for capturing and directing the fluid, evolutionary process of implementation."
>
> Thomas McNamara

How is implementation like seeing clouds? Taking off from this provocative analogy, Thomas McNamara, who directs Early Childhood Evaluation for the Philadelphia Public Schools, clarified some important similarities and differences.

The decision to "intervene in the contir.uous, massive movement of weather systems as they roll across the earth's surface" is made only when rain is considered absolutely essential to remedy drought conditions, McNamara explained. A comparable impulse to correct arid educational conditions prompts innovation in schools. The hit-or-miss characteristics of the art of weather control can be compared with the primitive state of the art of school intervention. And at present, ability to accurately predict the effect of educational programs is comparable to the state of weather predictions -- in both cases the knowledge-base is slim.

But here the analogy ends for, as McNamara pointed out, "the inanimate elements comprising the complicated web of interactions found in weather systems are far exceeded by the complexity of the self-knowing, abstract-thinking beings we have to deal with in the educational sphere." McNamara's long experience in schools has taught him that "complex human changes occur against a backdrop of existing human organizations" and cannot be reduced to low pressure systems meeting with high humidity conditions. Rather, the complexity of people's interactions within school settings must be understood in a "non-mechanistic, non-reductionist" framework.

What then is the most effective way to motivate school people to adopt new ideas and practices and to implement them with the commitment needed to affect change? According to McNamara, what is needed is a method that encourages healthy controversy and can lead to compromise and resolution. McNamara suggested the "Judicial Method" which relies on human testimony and enables people to develop a clearer understanding of the range of issues.

Specifically,

> "A trial held within the school-community context, would follow (with some modification) procedures of sound jurisprudential practice. There would be a 'judge,' a 'jury,' 'plaintiff,' and a 'respondent.' Witnesses would be called to testify in behalf of a position taken on one side or the other of a given issue. These witnesses would be examined and cross-examined as in a court of law. Pre-trial investigation would include interviewing a full range of potential witnesses. This investigation would also include the study and analysis of important documents, test scores, and other conventional assessment data to be presented later as exhibits during the public proceedings. The entire activity was envisioned as a clarification process ultimately leading not to a verdict but to a set of recommendations provided by a citizen jury. What was to be 'tried' was a range of important issues confronting the local school system. The guilt or innocence of persons within or without the system was not to be the issue. Indictment of individuals would serve only to subvert the major intention of the process—namely, clarification."[2]

"Since the very people affected by the emerging policy will be intimately involved in the inquiry process, the judicial method assures that the policy decisions are not only responsible but responsive to staff concerns," concluded McNamara.

[2] Wolf, Robert, "The Way I See It . . . American Education Should Go On Trial," (Educational Leadership, April 1980).

THE CONCERNS BASED ADOPTION MODEL

> "Outcome evaluations conducted after one year of use are apt to reflect less impact on students than perhaps even the previous year, when the innovation was not used."
>
> Susan F. Loucks, Gene E. Hall

A teacher adopting an innovation goes through six levels, according to Susan F. Loucks of the Network Inc., Andover, Mass.

These levels of increasing sophistication echo the well-known stages of intellectual and of moral development described by Piaget and Kohlberg respectively.

LEVELS OF USE OF THE INNOVATION: TYPICAL BEHAVIORS

LEVEL OF USE	BEHAVIORAL INDICES OF LEVEL
VI RENEWAL	The user is seeking more effective alternatives to the established use of the innovation.
V INTEGRATION	The user is making deliberate efforts to coordinate with others in using the innovation.
IVB REFINEMENT	The user is making changes to increase outcomes.
IVA ROUTINE	The user is making few or no changes and has an established pattern of use.
III MECHANICAL USE	The user is using the innovation in a poorly coordinated manner and is making user-oriented changes.
II PREPARATION	The user is preparing to use the innovation.
I ORIENTATION	The user is seeking out information about the innovation.
0 NON-USE	No action is being taken with respect to the innovation.

Loucks' and Hall's Concerns Based Adoption Model (CBAM) not only helps us understand the extent to which time and a series of adjustments in attitudes and skills are involved in change, it provides a framework to compare programs. If two models of innovation are competing for an administrators' favor, the one in which teachers reach the "refinement level" within two years might well be more attractive than the program in which most users never progress beyond the mechanical level. The Loucks and Hall analysis also suggests that evaluators would be wise to wait at least two years in any effort (when the results will reflect changed behavior that is more than merely mechanical), before judging a program's success.

Having a method to calibrate degrees of implementation has led Hall and Loucks to some additional discoveries. While it has long been thought incontrovertible that the more a program looks like its model, the better will be the results, they have found that often some degree of adaptation relates to higher outcomes than either high fidelity or major adaptations." They have also learned that often teachers must participate in the development, design and planning of the innovation, if they are to succeed with it.

These discoveries can explain a number of implementation oddities of the Planned Variation Follow Through experiment, particularly the repeated phenomenon (which became apparent at earlier NIE planning meetings) of school administrators singing the praises of certain "successful" models while at the same sites the model-sponsors bemoaned their failure because they saw local teachers deviating from the prescribed practice.

HOW CAN WE KNOW IF THE PROGRAM IS BEING USED?

"A first objective in future efforts to study program implementation . . . should be to provide an empirically based description of what the program is and is not."

Chad D. Ellett

Before a program can be evaluated, it is necessary to measure the degree to which it exists. One cannot appraise the effectiveness of an approach like Direct Instruction, for example, in a classroom in which the materials have been provided, but in which the teacher isn't actually using the program. So, the first order of business is to set criteria and collect data which will reveal the degree to which the program being evaluated is actually being conducted in a given classroom or school.

Chad D. Ellett of the University of Georgia, with Margaret Wang (University of Pittsburgh), have provided a framework for doing this. Briefly, their plan entails defining "critical dimensions" and "scaled descriptors" for each aspect of a given Follow Through Program. For example, one critical dimension would be the classroom teacher's communicating with learners by clarifying directions and explanations when pupils misunderstand. To measure whether a teacher was actually implementing this part of the program, an observer would apply a "scale of scoreable descriptors." On this particular scale the lowest rating might be "discourages learners when they seek clarification of directions," while a high scoring teacher would "give directions using different words when learners do not understand," and attempt to "identify areas of misunderstanding and restate communication."

Out of many such components the authors would construct comprehensive performance indicators for each critical dimension of Follow Through and a generic framework for evaluating program implementation. Such a framework for evaluation would, they conclude, "be 'generic' in terms of the *what* and the *how* of implementation, but flexible in nature in order to adaptively accommodate the diversity of Follow Through models" (emphasis in original).

RESEARCH NEEDS, SELECTION CONSIDERATIONS, AND ALTERNATIVE OUTCOME INDICATORS

> "Even if some agreement can be reached on the outcomes of interest, this does not guarantee agreement on the instruments to be used to measure the outcomes."
>
> J. Ward Keesling and
> Allen G. Smith

Whenever professionals in the field of program development or evaluation write or speak about the Follow Through program, they complain that only rather mundane measurement tools were used to evaluate the innovative Follow-Through models which brimmed with interesting and exciting consequences for children, parents, paraprofessionals, teachers, schools and communities. In these many-faceted programs:

- Children made leaps in learning, health and fitness, initiative, independence, and emotional growth;
- Parents learned to be firmer and more patient; to guide, to teach, to make personal decisions, and to collaborate with the schools;
- Paraprofessionals coped better in the marketplace and amassed credentials for new careers;
- Teachers and administrators developed new respect and understanding towards project children and their families; teachers' time and space management in class improved; administrators became more skilled in staff relations;
- The schools involved parents more, made necessary curricular changes, improved relationships with other educational agencies.

Yet these myriad outcomes were overlooked because acceptable results "funneled down" to a few narrow standardized measures of change.

J. Ward Keesling of Advanced Technology, Inc., and Allen G. Smith of System Development Corporation bemoan that waste. "Research conducted by the Follow Through sponsors themselves on alternative outcomes and measures covers at least 5 linear feet of shelf-space," recalled Keesling. "While many of the tests and measurements that were invented are too program specific and/or too expensive to administer widely, some of them could be useful in other programs."

The trick is to find the useful, inexpsensive, suitable material that matches the outcomes of a number of programs and make it available widely. Keesling and Smith's sensible proposal is "a review process similar to the Joint Dissemination Review Panel (JDRP) that would pass on the acceptability of the instrument for general use."

A FEDERAL ADMINISTRATOR'S PERSPECTIVE

> "Evaluations specifically designed to answer the questions ·sked by federal administrators may not help those who want to know what happened to the children."
>
> Garry L. McDaniels

People planning and implementing strategies to document school improvement efforts should know how federal programs like Follow Through are evaluated especially since they are likely to differ substantially from evaluations which would be of use to teachers, school administrators, and community leaders.

Garry L. McCaniels from the Institute for Program Evaluation, the Unitea States General Accounting Office, provided the following useful blueprint showing what questions standard evaluation of a Fede:al Program seeks to answer:

I. Identifying the goals of Congress

A. Who are the intended beneficiaries?

B. What services are envisioned for those beneficiaries?

C. What administrative mechanism did the Congress envision to provide services?

D. What positive impacts were expected? (What negative impacts were to be guarded against?)

II. Describing the executive branch'ɛ program

A. Who are the beneficiaries receiving services?

B. What array of services exist? What is the relative frequency of services?

C. What administrative actions has the executive branch taken? What administrative mechanisms are in place?

D. What impacts appear to be covered by or associated with the presence of these services?

III. Providing an analysis and syntheses of the data collected

A. Are the intended beneficiaries being served by this program? Are they receiving services they might not have been otherwise receiving as a result of this program?

B. Are the services being received consistent with those envisioned in the Act?

C. Are the actions taken by the executive branch consistent with those expected by the Congress (e.g., regulations, distribution of effort)?

D. Are the impacts identified related to the services provided and are these impacts consistent with the intent of the legislation?

IV. Providing recommendations

A. for the law

B. for the executive branch

C. for the local administration of services and/or federal funds.

The immensity and complexity of this agenda of questions has brought McDaniels to the conviction that the job is one for many hands.

"Experience has led me to believe," said McDaniels, "that it is physically and intellectually impossible for a single organization to organize and execute a major program evaluation because no single organization has a sufficient pool of talent, and techniques favored by researchers working for a particular organization tend to favor similar techniques of data gathering. As a result, when I hear that one RFP has been announced to evaluate a given program I feel the policy-maker will be poorly served."

McDaniels would like federal agencies to use several different contractors -- each chosen because they can contribute uniquely to an aspect of the program being studied. He would also like to see a number of different methodologies employed -- each specifically designed for a specific question. But no one investigator should be made to feel that his or her study should have the goal of clarifying all aspects of a major policy question.

Finally, McDaniels feels that a final report should be commissioned to synthesize all the studies and to clearly identify the cumulative meaning of findings. The individual reports of investigators should not leave out important details for the sake of "untechnical" readers.

"The investigator reports for a federal evaluation should be good examples of scientific writing -- technically responsible and readable," he concluded.

ASSESSING LANGUAGE-MINORITY STUDENTS

> "Hispanics and other language minority groups have become victims of test abuse and test misuse."
>
> Ernest M. Bernal

"The only group that profits from the use of English-based achievement tests on limited English proficient children are the test makers," contended Ernest Bernal of Creative Educational Enterprises (Austin, Texas). He argued that the tests are harmful to the children and of little use to educators. Unless they are redesigned, "most of the achievement and affective data will be worthless."

According to Bernal, the present tests are unreliable, except for the short run. When language-minority children are tested in English, the tests inadequately assess their aptitudes, attitudes, achievement and development. Nor do they predict which students are likely to succeed.

While the test results therefore have been of little practical value, they have had considerable negative effect because teachers often predict childrens' failure on the basis of test results and give up on them.

Bernal reminded us that because these children (singly, and in a group) are different, if their scores are to be included in the new Follow Through program evaluations, student variables unique to this group will have to be dealt with, i.e.:

- their competence or lack of it in both English and their own language,
- their general communicative competency,
- their achievement in select subject areas,

- their cognitive style,
- their self-esteem, inter-ethnic attitudes and own-language attitudes.

These children also test differently, which makes them hard to evaluate, explained Bernal. No one has succeeded in correctly interpreting the test scores of students who are taking math or science achievement tests and are not proficient in the language of the test. Sometimes the results are startling -- as when children show a sudden extraordinary pre-test to post-test gain (which merely means that between the tests, they have learned to read). On the other hand, the scores of those who don't learn to read get worse as the norm expectations increase in difficulty. So when the scores of these two groups are averaged, "Presto! No gains!" Bernal asserted that during the evaluation of the first Follow Through programs, some evaluators were so stumped that they "pulled" their scores so that they wouldn't be included in the analysis.

In addition to student variables, variables specific to ESL and bilingual programs inevitably confuse evaluation results. We need to know:

- the language proficiency (oral and written) in both languages of teachers and aides.
- the proportions of instructional time and content in English and the non-English language.
- which of the instructors (the more prestigious teacher or the less prestigious aide) conducts instruction in English and which in the second language.
- the type of bilingual or ESL instruction provided.

"If we do not consider deliberately the relationship of language minority students to the entire new Follow Through effort, their presence by design or accident may become a nuisance, a 'noise' or cacophony which our interventions, instruments and methodology are ill-prepared to orchestrate," warned Bernal.

EXPANDING THE USES OF TESTS

"If we view standardized tests not simply as measurement instruments but as sources of direct learning, then perhaps we might develop them in different ways."

Walter Haney

For years tests have been used to sort children (I.Q. tests), to uphold educational standards as antidctes to grade inflation (Regents and achievement tests), and as debating material in a continuing discussion on the main aims of education (high school competency tests). But we have not yet found ways to construct tests that are terribly helpful as direct aids in teaching and learning, according to Walter Haney of the Huron Institute.

Norm-referenced tests are unsuitable for measuring a program's effectiveness because they are constructed to be insensitive to the effects of instruction in local school systems (which all have different curricular). Nevertheless they are now frequently used for this purpose although their results can be misleading, said Haney. Even the criterion-referenced instruments which are designed to measure a programs' effectiveness are not sufficiently refined to do this well. "Work on criterion-referenced measurements seems to be progressing far faster on technical issues, such as methods of item-analysis, setting cut-off scores, assessing decision consistency, and applying generalizability theory to analyze variance in test results than on the substance and skills of what has or has not been learned."

Haney would like to see more effective tests for program evaluation, and a new emphasis in test-making -- tests that actually teach children.

Would the primary function of tests -- the sorting and policy making functions -- be violated by developing tests that aid instruction? Haney thinks not, pointing out that the uses of tests have changed through the years and will no doubt change again. "Not many years ago, educational program evaluation was viewed as research in the service of de:ision-making, but studies since then have shown that findings rarely have contributed directly to decision-making in the way that was expected. Now program evaluation is seen less as applied science and more as a descriptive enterprise, and it is possible that testing as part of the evaluative enterprise could be aimed less at formal inference and selection and more at description."

How would tests be developed from which people could learn? Haney thinks a reasonable place to begin would be with theories of learning such as Benjamin Bloom's theory of mastery learning which highlights four elements of "quality instruction": cues, participation, reinforcement and feedback.

If tests were to be designed as learning instruments they might provide:

1. Cues that could be altered or adapted to present those which work best for particular learners -- i.e., written cues for some students, oral cues for others.

2. Opportunities for active participation and practice with differences in the amount of practice or participation depending on the individual learning style and needs of students.

3. Reinforcers which would be adapted to the particular learner (since what is a reward for one child may not be for another).

4. Quick and corrective feedback for students, when and where needed.

"When tests are viewed strictly as measurements, alternative modes might be viewed as a problem, namely as extraneous sources of error variance. But from the learning perspective, alternative modes might be viewed more positively as differentially appropriate for students with different learning styles," says Haney.

"Specifically they might:

- be available in alternative modes of presentation e.g., oral, written and via video screen rather than simply written

- be labeled in terms familiar to test-takers rather than in terms of psychological constructs on behavioral domains (e.g., word wizard tests rather than vocabulary tests)
- be self-scoring or scoreable by individual test-takers
- be of variable length
- provide results not only on whether answers are right or wrong but on the nature of errors or sources of corrective instruction."

Haney concluded by suggesting that the role of evaluators could be changed from "producing knowledge to give to educators for purposes of educational improvement" to "providing tools to educators and society generally with which to communicate about education goals and values and providing instruments to learners to improve learning."

ASSESSING ABILITIES OF BLACK CHILDREN

"Test items should be designed to elicit the most sophisticated, complex or at least the most appropriate cognitive processes in these children."

Dalton Miller-Jones

Alice has been prodded and prompted by adults since birth. She has learned to "read" adult questions. Whenever her mother or grandmother asked "What kind of fruit do you want, an apple? An orange? A banana?" and Alice answered, "A cookie," she was gently reminded they said fruit. By three, Alice wasn't making that "mistake" any more. She knew what they wanted to hear. She learned to say, "I don't want a fruit, I want a cookie."

Betty's mother worked. When she wanted something to eat, she knew where to find it, and got it for herself. Often her older sister simply shared what she was having without asking.

So, as Dalton Miller-Jones from the University of Massachusetts pointed out, while both children know that apples and oranges are fruits, Alice will always say they are similar because they are fruits, Betty may tell you that what is most important about their similarity to her is that she likes eating both of them. Alice is said to have a higher I.Q. because she knows what adults want to hear.

Charles is blond. When asked what a brunette is, he says it's a person with "dark brownish hair." Daryl is black, he tells you a brunette has "light brownish hair." Although the dictionary says brunette is "a reddish moderate brown," Charles is "right" while Daryl is "wrong." Since this is a question from an I.Q. test, Charles' I.Q. mark is higher than Daryl's.

These and other such questions used as indicators of "intelligence" statistically "prove" black children have lower intellectual ability than white children.

According to Miller-Jones, the logical inconsistencies in the standardized I.Q. tests are legion. It is correct to say that houses are made of bricks and wood but incorrect to say they are made of sticks and nails. It is correct to say windows are made of "glass and wood" but incorrect to say they are made of "screens and putty." It is correct to say books are made of paper, plastic and something hard for covers but incorrect to say they are made of "pictures and pages."

As Miller-Jones pointed out, "there appears to be no intellectual distinction (by test makers) between acceptable and unacceptable responses -- and there is no consistency in the criteria invoked . . ." Perhaps of more consequence, there is no feedback to the child. Children who are not trained as was Alice, to know from experience what adults want, will answer the first thing that comes to mind and assume, because they get no negative feedback, that any answer is as good as any other.

Without this feedback, black children are likely to give unacceptable answers. It is also argued that minority children have different cognitive styles and cultural traditions developed as a result of their different environments. These children are affectively oriented and use what could be considered relational styles while schools typically support and are oriented to analytic styles. Miller-Jones concurs with Asa Hilliard of Georgia State University that unlike Euro-Americans who tend to believe that anything can be divided and subdivided into parts and these add up to a whole, "Afro-Americans tend to respond to things in terms of whole picture instead of its parts; that [they] prefer to focus on people and their activities rather than things or objects; that [they] lean toward altruism and social cooperation; and that [they] tend not to be 'word' dependent for meaning, relying heavily on actual behavior and experience."

Citing further evidence of differences in cognitive style provided by other researchers, Miller-Jones suggested that these children probably need more varied stimululi for learning and more practice in the "accepted" modes of analysis. To assess the language and cognitive development of these children, Jones recommends multiple cognitive and language eliciting materials, culturally salient subject matter and materials, a familiar comfortable test environment, a variety of tasks relative to inductive rather than deductive styles, and a demonstration of what is expected through concrete examples given before the test.

Miller-Jones believes more research is needed in determining the relation of conceptual and cognitive styles to school performance in reading, math and social studies. Equally useful would be diagnostic profiles which probe how students arrive at answers. "What are we asking children to do that sends some children down blind alleys to pursue unfruitful strategies?" asked Miller-Jones. "We don't know now, but we do know that some children latch on to 'good ideas' that work for a while but ultimately subvert their learning process like sticking with first letters and context when reading, or memorizing books until their memories give out. This is not just an issue of minority assessment," he continued. "To help all children fulfill their potential we have to study the cognitive operations of the children who seem to get it right all the time, along with those who persist in getting it wrong."

DISCUSSION

Following the presentations at each of the first three sessions a lively discussion led by pre-selected experts around the table provoked the speakers to clarify their thinking and defend their positions. This discussion is reflected in the synthesis of each of the presentations.

Probably the liveliest interchange followed the presentation of the papers of Glass, House, Burstein and McNamara. There was basic agreement that during the first sixteen years of Follow Follow Through greatstrides had been taken--but the large scale evaluation studies were disappointing. It was the dimensions of the disappointment and more important, the reasons for the disappointment that provoked controversy.

Some agreed with Edward Zigler from Yale, one of the prime designers of Follow Through, who felt it was a good experiment badly executed. Others sided with David Weikart from High/Scope that the problem of Follow Through was its false presumption that we could find the one best method of educating all children. Weikart, who also took part in the 1968 planning meeting of Follow Through, reminisced that this was another round of the same argument that dominated discussion back then -- between those who sought the ideal answer for everyone and wanted the whole educational establishment to embrace it once it was found, versus those who felt the necessity to keep going back to check the social context and what the individuals within it need.

Ernest House put the argument in an even broader historical perspective, tracing back to the Renaissance our mistaken belief that the scientific method could provide suitable answers to all questions. He proposed a counter notion: humanistic inquiry.

The discussion did not resolve the philosophical differences around the table. But by the time the session was curtailed by the demands of the schedule, the sides were clearly drawn, with Edward Zigler regretting the "bad science" that created the evaluation problems of Follow Through, and Ernest House responding that even if we did it again well, we'd still have a mess because it was built upon false presumptions.

The most comprehensive commentary on the implementation presentations was delivered by Convener Wang. Putting the concerns of the papers into a larger perspective, she arrived at three basic sets of recommendations for NIE and the field's consideration:

"1. Models neither have unified implementation or effects across sites, nor do they replicate easily or in similar processes from one site to another. Therefore, to continue the pattern of trying to identify which educational approach is best for disadvantaged children, even if program implementation variables are included in the evaluation design, is not only unproductive but will tend to yield misleading evaluations.

2. Implementation of innovative school improvement programs continues to change. The implementation process is affected not only by the nature of the intervention but also by a host of factors that vary from situation to situation. Therefore, evaluation of innovative programs requires a developmental perspective and dictates the use of a longitudinal design with repeated measurement focus on 'improvements' rather than 'proof.'

3. The study of implementation requires an interactive and multifaceted approach using multiple criteria and methods of data collection and analysis. Information is needed not only for use by the consumers of the innovations to improve their program implementation but also to further our understanding of

the implementation process. Such information can facil :ate the widespread adoption of innovations for meeting school improvement needs in a variety of school contexts."

CONCLUSIONS AND RECOMMENDATIONS

During the fourth session four groups made up of the conference participants met simultaneously to make summary recommendations to the NIE and to the field at large. Because the participants in the discussion groups are distinguished leaders in the field, we have listed them by group to give the reader a sense of the philosophical and clinical mix of those who concurred in the final recommendations.

Group I: FIRST THINGS FIRST

Topic: Supporting research for the evaluation of NIE-fu.ded pilot Follow Through projects

Marianne Amerel, Leigh Burstain, Celestino Fernandez, Ernest House, Lawrence Rudner, William Tikunoff, David Weikart.

Group II: RESEARCH TO IMPROVE THE ART OF RESEARCH

Topic: Research related to methodological and technical developments in program evaluation

William Cooley, Chad Ellett, Hortense Jones, Tom McNamara.

Group III: TOWARDS NEW TESTS AND BETTER TESTS

Topic: Supporting research on instrumentation and development of program implementation and outcome measures

Ernesto Bernal, Edmund Gordon, Walter Haney, J. Ward Keesling, Susan Loucks.

Group IV: THE NEXT ORDER OF BUSINESS FOR COMPENSATORY EDUCATION -- R&D

Topic: Issues and agenda for research and development in compensatory education

Freda Holly, Dalton Miller-Jones, Garry McDaniels, Eugene Ramp, Margaret Wang, Edward Zigler.

Group I: FIRST THINGS FIRST

The group focusing on supporting research needed for the evaluation of the new NIE-funded pilot projects made three major points regarding the content, the methodology, and the dissemination of evaluation.

Content:

An evaluation report makes a good deal more sense in context. If it is accompanied by a detailed portrait of the school and the community, readers can understand how the local culture and economics shaped the program and what effects the program has had on local conditions. Therefore it is helpful to collect uniform descriptive data even before the program starts. This data can help:

- researchers and evaluators to understand the results in context
- administrators from outside the district to make knowledgeable decisions when they consider adoptions
- administrators from inside the district to compare their outcomes with outcomes from districts with similar demographic and economic conditions.

Methodology:

Evaluation modes are needed that can serve a number of programs and models. But a distinction must be made between evaluation data that covers a common ground (as described by Ellett in his presentation), and evaluation data that uses precisely the same tests to measure the effectiveness of models, as was done in the first round of Follow Through evaluations with unfortunate results. In that attempt to compare the effectiveness of programs, all programs regardless of their intents were submitted to the same tests. The process was criticized by a large number of programs that were judged to have lost a race they hadn't tried to enter.

Dissemination:

The language of evaluation must be refined. Evaluations should not only communicate to sophisticated administrators and other evaluators, but to parents, teachers and lay boards of education.

A final suggestion was not related to evaluation, but to the benefits of retaining the sponsor-site structure of Follow Through. The group urged that the sponsorship of programs has worked well in the past and should be continued. Sponsorship has:

- linked local schools with agents outside the school who could give them training, support and guidance
- helpec locals through difficult implementation problems by drawinq on the experience sponsors amassed facing similar problems in other locales
- helped protect services which addressed the needs of children when they were threatened by local political considerations.

But although the experienced sponsors have a track record, this should not preclude new sponsors from being invited to respond to the upcoming Requests for Proposals, as long as they too are required to specify their models' instructional intent and demonstrate prior experience in implementing educational innovations.

Group II: RESEARCH TO IMPROVE THE ART OF RESEARCH

Needed developments in technology and methodology were the focus of one discussion group which proposed that the new NIE-funded Follow Through research program should be viewed, in part, as a "laboratory" for conducting more detailed studies of how evaluation as a process could promote and facilitate school improvement. Computers, which have thus far been under-utilized in the field, should be more prominently used. They can be used by sponsors and local school districts who want the latest evaluation information and test designs. As J. Ward Keesling pointed out in his presentation, many programs could use an "item-bank" of outcome measures that would be closely enough matched to the intent of each program to be meaningful, yet generic enough to allow for comparison across programs. Computers could be the best way to make these available to the field and keep the information current.

But while new Follow Through projects should advance the art and science of evaluation, that should not be the evaluators' central focus, the group further urged. Nor should evaluators on-site merely document the final success or failure of a project. Rather they should employ their skills to help solve the obvious problems facing the district, and even locate subtler problems that inhibit program implementation and accurate assessment.

For example, tests are not yet perfected that are sensitive to minority and low income children. Such measures would be helpful to local school districts to better evaluate their individual school improvement efforts. At the moment they have only nationally normed instruments to work with, which sometimes mask the great strides they are making.

To be accurate and useful the evaluations should not merely measure academic achievement, the group concluded, but should address head-on the greater challenge of measuring those more complex effects which were the programs' original concern (also see group IV). The evidence points to the need to balance data collection with descriptive narrative, as suggested by Ernest House.

Group III: TOWARDS NEW TESTS AND BETTER TESTS

Four approaches were urged to improve the instrumentation available:

- the expansion of what testing can do
- the use of tests to find more refined ways to reach and teach children
- a consortium to maximize the efforts of the educational community in developing instrumentation and techniques
- the development of measures of program implementation.

Tests That Teach: While we need new generalizable outcome measures that cover more ground than do standardized achievement tests (see groups I and II), some time and energy should be spent devising tests that are educative devices for children -- tests that teach. And we also need tests that are designed to shed light on children's learning styles or skill mastery when "read" by teachers trained to use the test data (as described by Haney).

Test Administration: Researchers, as Miller-Jones and Bernal pointed out, have demonstrated that test directions and administration favor children who understand the implicit rules and penalize those who don't. Why? One reason is that we do not know enough about how to administer tests, or give directions that will be fully understood by all children. And we do not know what effect the test-taking environment has on some children. Additional research is also needed to better understand the many different ways children arrive at answers to tests as a possible way to improve their approach to cognitive problems in both test and learning situations.

A Consortium Approach: Finding the time, money, and expertise to improve, refine and invent tests as discussed above would be exceedingly difficult for single sponsors or schools to undertake. So the group envisioned a consortium of sponsors, academics, teachers, parents and local educators that could be managed and assisted by the National Institute of Education to develop new measurement tools.

Measuring Implementation: Echoing the concerns expressed the prior day by Ellett and Loucks regarding the need to pinpoint levels of implementation at sites (see groups II and IV), this group recommended the development of tools for this purpose. Assuming that certain programs are more difficult than others to convey to teachers, researchers should explore their cost-effectiveness from the point of view of how long it takes before innovative programs are actually put into practice. The narrative-descriptive report-format (described by House) was considered to be especially appropriate here.

One member of the group suggested that perhaps a more seminal question -- whether supporting research ever actually results in changes in school practice -- would be worth pursuing.

Group IV: THE NEXT ORDER OF BUSINESS FOR COMPENSATOPY EDUCATION -- R&D

The group which set out to develop an agenda for research and development in compensatory education spotlighted three areas of concern: what happens to children, how to maximize the usefulness of tests, and how to facilitate school change.

1. The first prioritv should be to learn what happens to children when they move from one situation to another:

   - from one grade with a distinctive program, to the next with a different one
   - from home with one set of expectations, to class with another
   - from the communitv with one dominant culture, to the school with another.

Researchers have explained how cognitive styles of minority children frequently impede them in school. But we know little about what problems are created by cultural differences. The new round of contracts should measure a broader set of program results than did the last Follow Through evaluations which focused on academic achievements (a point made also by groups I and II). But to do so adequately will require refinement of techniques to assess childrens' progress, describe their program experiences and find the means to measure what is accomplished by programs that stress process over product.

2. Current testing programs should be studied to improve and expand the use of tests. Specifically the group wants to:

   - maximize information obtained from tests

- discover what various test responses say about the child and how the insights gained can be used to help that child succeed in school

- close the gap between what teachers teach and what tests test.

3. Finally, Follow Through should serve as a national laboratory for studying schooling in grades one through three. The nationwide agenda should include investigating deterrents to innovation and change in the nation's schools; describing how program implementation is accomplished in terms school people can comprehend (a point made also by groups II and III); monitoring the utility and accessibility of the literature on innovation; and identifying the most promising aspects of the partnership between parents and teachers.

AFTERWORD

Transcending the specific recommendations and conclusions reported above, the conference generated a spirit of commitment best expressed by convener Margaret Wang:

> "Research designed to produce useful information for school improvement is no longer just an ideal. It is becoming a reality, largely through the kinds of capabilities and technological advancements in programs evaluation discussed here. But continued progress depends on scholarly advances in research aimed at building our knowledge of what is being implemented in our schools and how. Supporting such would be a fruitful investment of further public funds . . ."

APPENDIX

CONFERENCE PARTICIPANTS

Marianne Amarel
Educational Testing Service
Princeton, NJ 08541

Ernest Bernal
Creative Educational Enterprises
5203 Hedgewood
Austin, TX 78745

Jacqueline Blackwell
Indiana University
School of Education
902 N. Meridian Street
Indianapolis, IN 46204

Gerand Burns
Dept. of Education
400 Maryland Avenue, S.W.
Transpoint Bldg B-432
Washington, D.C. 20202

Leigh Burstein
Dept. of Education
University of California
Los Angeles, CA 90024

William Cooley, LRDC
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Chad D. Ellett
College of Education
University of Georgia
Athens, GA 30502

Celestino Fernandez
Dept. of Sociology
University of Arizona
Tucson, AZ 85721

Robert Glaser, LRDC
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Gene Glass
Laboratory for Educational
Research
University of Colorado
Boulder, CO 80302

Edmund Gordon
Dept. of Psychology
Yale University
New Haven, CT 06520

Walter Haney
Huron Institute
123 Mt. Auburn Street
Cambridqe, MA 02139

Freda Holley
Austin Independent School
System
6100 Guadalupe
Austin, TX 78752

Ernest House, CIRCE
University of Illinois
270 Education Bldg
Urbana, IL 61801

Dalton Miller-Jones
Dept. of Psychology
University of Massachusetts
Amherst, MA 01003

Hortense Jones
New York City School
131 Livingston Street
Brooklyn, NY 11201

J. Ward Keesling
Systems Development Corp.
2500 Colorado Avenue
Santa Monica, CA 90406

Susan Loucks
The Network Inc.
290 South Main St.
Andover, MA 01810

Garry McDaniels
U.S. General Accounting Office
441 G Street, N.W.
Washington, D.C. 20548

Thomas McNamara
The School District of
Philadelphia
21st Street-The Parkway
Philadelphia, PA 19103

Eugene Ramp
Dept. of Human Development
University of Kansas
Lawrence, KA 66144

John Rodriguez
Acting Assistant Secretary
for Elementary and Secondary Education
400 Maryland Avenue, S.W.
Washington, D.C. 20202

Walter Stalford
National Urban League
500 East 62nd Street
New York, NY 10021

William Tikunoff
Far West Laboratory
1855 Folson Street
San Francisco, CA 94103

Margaret Wanq, LRDC
Universitv of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

David Weikart
High/Scope Education
Research Foundation
600 North River Street
Ypsilanti, MI 48197

Joseph Wholey (USC)
Washington Public Affairs Ctr.
512 10th Street, N.W.
Washington, D.C. 20004

Edward Zigler
Yale University
Dept. of Psychology
New Haven, CT 06520

Conference Synthesizers

Beatrice Gross
Ronald Gross
17 Myrtle Drive
Great Neck, NY 11021

NIE Representative

Lawrence M. Rudner, Ph.D.
Senior Associate
National Institute of Education Room 822
1200 19th Street, N.W.
Washington, D.C. 20208

PAPERS

Bernal, Ernest M., "Assessing Language Minority Students in the New Follow Through."

Burstein, Leigh, "Methodology for Evaluating Social Programs When Individuals Belong to a Variety of Groups Over Time: Implications for Follow Through Research and Evaluation."

Ellett, Chad D., "Issues Related to the Evaluation of Program Implementation in Follow Through."

Glass, Gene V., & Camilli, Gregory A., "'Follow Through' Evaluation."

Haney, Walter, "Thinking About Test Development."

House, Ernest R., "Scientific and Humanistic Evaluations of Follow Through."

Keesling, J. Ward, & Smith, Allen G., "Issues Related to Instrumentation in Large-Scale Program Evaluation: Research Needs, Selection Considerations and Alternative Outcome Indicators."

Loucks, Susan F., & Hall, Gene E., "Investigating Program Implementation: A Field Perspective."

McDaniels, Garry L., "A Federal Administrator's Perspectives on the Documentation of School Improvement Efforts."

McNamara, Thomas C., "Charting the Course of Implementation."

Miller-Jones, Dalton, "Future Follow Through Documentation and Research: The Assessment of Academic/Cognitive Abilities of Black Children."