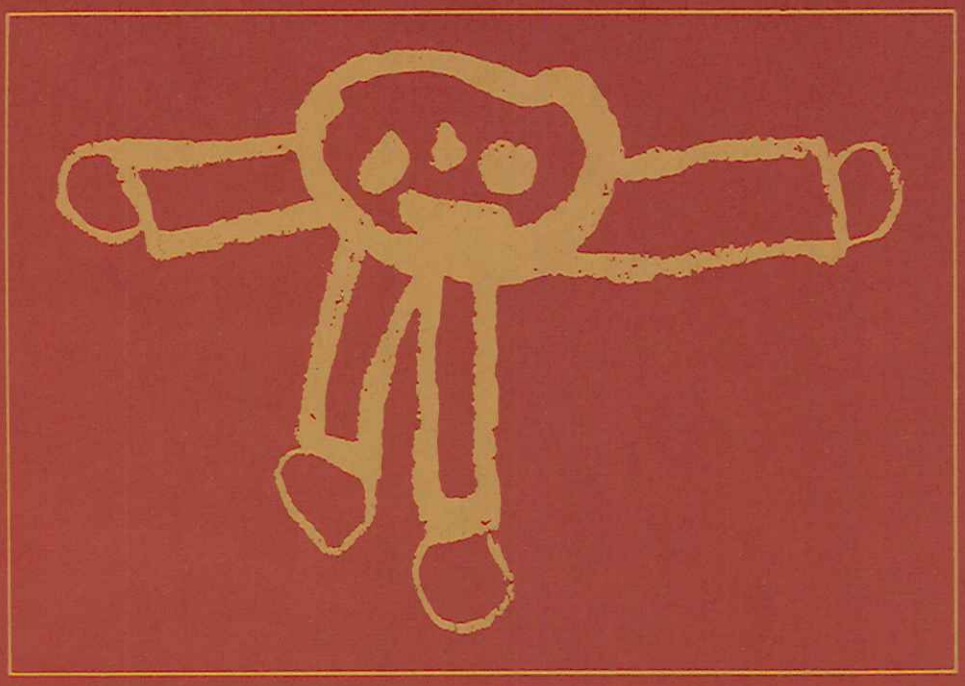
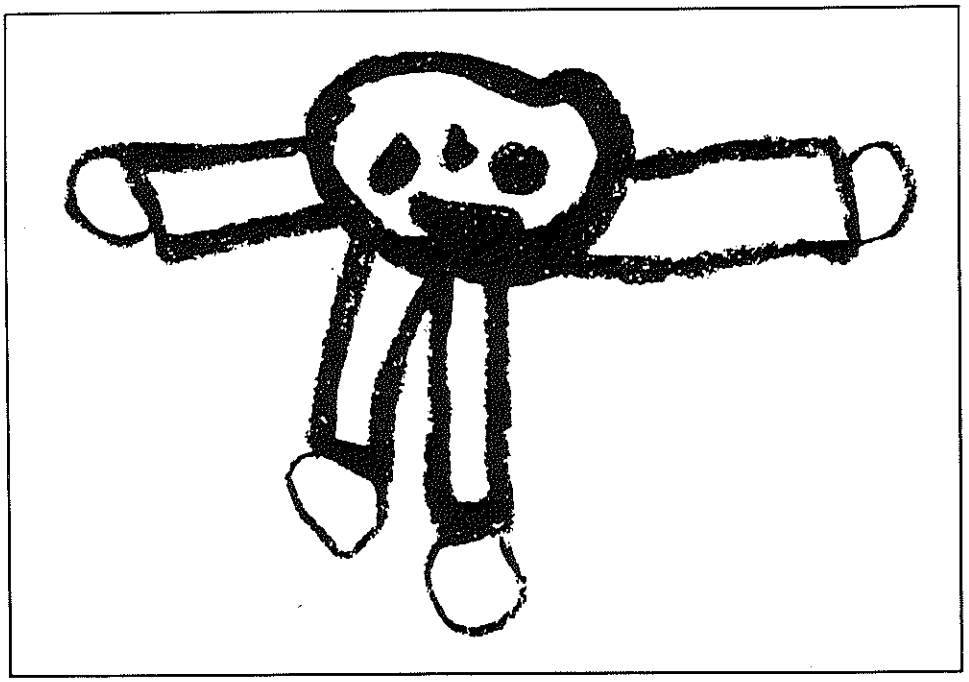
North Dakota Study Group on Evaluation



George E. Hein

AN OPEN EDUCATION PERSPECTIVE ON EVALUATION

North Dakota Study Group on Evaluation



George E. Hein

AN OPEN EDUCATION PERSPECTIVE ON EVALUATION

University of North Dakota
Grand Forks, N.D. 58202
February 1975

Copyright © 1975 by George E. Hein

First published in 1975

North Dakota Study Group
on Evaluation, c/o Vito Perrone,
Center for Teaching & Learning
University of North Dakota
Grand Forks, N.D. 58201

Library of Congress Catalogue
Card Number: 75-277

Printed by University of
North Dakota Press

A grant from the Rockefeller Brothers Fund
makes possible publication of this series

Editor: Arthur Tobier

In November 1972, educators from several parts of the United States met at the University of North Dakota to discuss some common concerns about the narrow accountability ethos that had begun to dominate schools and to share what many believed to be more sensible means of both documenting and assessing children's learning. Subsequent meetings, much sharing of evaluation information, and financial and moral support from the Rockefeller Brothers Fund have all contributed to keeping together what is now called the North Dakota Study Group on Evaluation. A major goal of the Study Group, beyond support for individual participants and programs, is to provide materials for teachers, parents, school administrators and governmental decision-makers (within State Education Agencies and the U.S. Office of Education) that might encourage re-examination of a range of evaluation issues and perspectives about schools and schooling.

Towards this end, the Study Group has initiated a continuing series of monographs, of which this paper is one. Over time, the series will include material on, among other things, children's thinking, children's language, teacher support systems, inservice training, the school's relationship to the larger community. The intent is that these papers be taken not as final statements--a new ideology, but as working papers, written by people who are acting on, not just thinking about, these problems, whose implications need an active and considered response.

Vito Perrone, Dean
Center for Teaching & Learning,
University of North Dakota

Contents

The widespread use of tests for purposes of selection, for deciding from Kindergarten on up who will pass and who will fail, who will be winners and who will be losers, is not likely to go away in a hurry. For, whether we like it or not, it has become indigenous to the kind of competitive culture that characterizes our social institutions, including our educational institutions. *Henry S. Dyer*

Introduction

George Hein is former director of Follow Through at Education Development Center, Newton, Mass., and during the current academic semester is coordinator of independent studies at Lesley College in Boston.

The common practice of assigning the task of making judgments about programs or children's achievements to outsiders stems from a desire that evaluation be 'objective', exempt from local influences, and applicable to any number of different situations. This concern with objectivity can be traced to efforts in the social sciences to attain the objectivity of the physical and biological sciences, where, according to popular belief, a description of an experiment by one group should be such that any other competent scientist can repeat the experiment; where the description of what happened should be so divorced from local events, or time-dependent parameters (except for events in time), that any other competent person could repeat them.

Although such a position is widely accepted in the physical sciences as appropriate *in principle*, working scientists know that many times they cannot repeat other experiments, or that it simply is not worth the bother to make the effort to duplicate them. What is required in physical science work is: first, that the results reported be consistent with accepted theory and, secondly, that the results reported or the compounds prepared show the properties that one would expect of such materials in the common course of events.

In evaluating educational work, the public can also expect, first, that the evaluation effort be consistent with the practice assessed: that it show reasonable results or ascribe properties to the educational system or outcomes consistent with what is generally known about children and learning; and, second, that the description of the evaluation activities and their relation to the program are such that other interested parties can carry out similar activities or compare them with their own experience.

It is naive, however, to assume that an evaluation is objective simply because it is carried out by someone not connected with the ongoing activity. Certainly people with a stake in a program must work out some way to recognize their own enthusiasm and self-interest in what happens and therefore take measures to minimize this influence. But a realistic recognition of this problem is more likely to result in relatively objective evaluations than does the reliance on outsiders. 'Professional' outside observers are still human and therefore as open to the

Support for the preparation of this paper was provided by a Study and Travel Grant from the Ford Foundation.

I gratefully acknowledge the constructive criticism of Ruth Ann Aldrich, Brenda Engel, Deborah Meier, Emily Romney, David Rubin, Frank Speizer, Jean Speizer, Ruth Schmitt and Lillian Weber in the preparation of this manuscript. *G.E.H.*

same problems of honesty and objectivity as anyone else.

But the problems surrounding evaluation are greater than any methodological issue. Each kind of educational philosophy requires its own approach to evaluation. An analysis of evaluation has to ask what the goals of the program are, and how any evaluation strategy supports and influences the program. In other words, evaluation must be considered both for itself and for its impact on the total program, not as a separate activity carried on outside the confines of the rest of school. Like curriculum, teacher training, and school organization, evaluation activities are an integral part of school, influencing every other part.

1

Open Education Principles Relevant to Evaluation

I can establish some sense of the evaluation and measurement strategies suitable to open education by listing briefly some of the principles on which the open education movement is based.

DEVELOPMENTAL ISSUES

First and foremost, open education includes the belief that the individual growing child is educable and stands at the center of the educational process. This data has, of course, been the rallying cry of educational thinkers in the liberal tradition for hundreds of years. The statement takes on new meaning, however, when current knowledge of children's growth and development is added to it.

It is now recognized not only that children are different from adults, but that children differ from each other; they go through developmental stages at varying rates and with varying learning styles. Child development experts no longer speak about '*the* six-year-old', but about the range of activities, ideas and concepts which different individuals in a group of six-year-olds will exhibit. Educators and parents have to ask whether the evaluation programs currently available in our schools take into account this variability in development.

Another difference in learning style among children is in the 'horizontal' dimensions of their growth and development (Bussis and Chittenden, 1973). Individual children not only reach different stages of development at various times in their lives, they also need to spend various amounts of time confirming and internalizing those stages. Most people know the distinction between learning the meaning of a new word and being able to use it unself-consciously in speech. There is always some span of time when one tentatively tries out the word, perhaps plays with it a little, listens to the sound in a sentence, and sees its effect on others before one can safely and naturally use a new expression. This learning time will vary from word to word, from situation to situation, and according to the need to use the word. If one watches children learn to speak, this phenomenon is apparent. The same pattern occurs, of course, with all the concepts and ideas anyone acquires.

Horizontal dimensions of learning do not show up as a mastery of greater numbers of words or knowledge of more concepts. Instead they manifest themselves in the richness of associations that a child is capable of, in the variety of ways that a concept or word can be used. They are an important part of the learning process, although seldom measured in evaluation procedures. Any educational program that takes individual children seriously has to take this horizontal component of learning into account.

Along with the recognition of children's individual rates and styles of growth comes a reluctance to put rigid timetables on the acquisition of skills or knowledge. In recent years a few psychologists have argued that unless children learn to read by the age of six or seven or acquire other school-related skills in the early years, they will be permanently behind. (Much of this argument stems from the 'cultural deprivation' perspective on poverty, which claims that the difference between poor and rich is that the 'culture of poverty' does not include the appropriate formative experiences children need to benefit from school.) But this view has recently been put into question by findings of societies where young children are kept very quiet and inactive yet grow into active intelligent adults (Kagan, 1972). Given such divergent evidence, it is more productive to look at the variation among individual children, to study their styles and their growth, than it is to try to generalize about the maximum necessary conditions for rapid attainment of skills and accomplishments.

Finally, advocates of open education believe strongly that the way in which children can progress through the various stages of development to more adult understandings of the world around them is through exposure to that world. Like good physical growth, which is only partly determined and requires nourishing food and regular use of muscles and limbs, mental, emotional, and social growth requires constant, active involvement with the rest of one's immediate world. Children's understanding of the physical world comes about as they play with things, observe them, manipulate them and generally begin to affect them. Children also learn about themselves and other people, about feelings, about cooperation or competition, by being in social situations and exploring their ramifications. This interaction with the things in the world is not only important for learning about these things, it is essential if children are to learn *how* to learn about them.

SOCIAL ISSUES

Another set of beliefs about children, social in origin, makes the findings of developmental psychology especially significant. A belief in the educability of all children, and the recognition of the individual qualities of each child, is essentially a belief in the *value* of each indi-

vidual--child or adult. It follows then that if we respect each individual, we must be concerned with his or her personal growth.

To provide maximum opportunities for each individual to grow and develop most fully, it is necessary to minimize the social influences that prevent the attainment of these goals, and to do everything possible to avoid damaging or stifling situations. If there is a question of nutrition or physical health, educators who are concerned for the growth of individuals will do all they can to see that children are well fed and healthy. Likewise educators must also combat social diseases that threaten to harm the individual child: forms of prejudice or stereotyping that force children into roles or categorize them independent of their individual qualities. Racism and sexism, among the most intense forces in our society, or any practice or tendency that categorizes children arbitrarily by some external factor, robs them of some portion of their ability to grow and to learn. Stereotyping gets in the way of seeing a child as an individual, interferes with providing experiences for a child from the whole range of life, and diminishes the opportunity to follow an individual timetable of horizontal and vertical growth.

Once children are placed in categories according to some specific quality they evidence at a particular time, there is an inevitable ordering of those categories and a decrease in the respect which is shown to those children during their development. Two major components of this trend can be ascertained. First, there is an inevitable ranking of children by the style they show. The most common practice in schools, which invariably places children into categories judged in an order from 'best' to 'worst', is tracking--the setting up of streams for the children on the basis of the rate at which they attain certain skills. I know of no system of tracking which is not also accompanied by a value judgment about which are the best children and which are the worst. Yet even the slightest knowledge of developmental theory should tell us that how fast one learns something has very little to do with how well one learns it or how much one can learn. The almost universal outcome of sorting children is that they stay in the group into which they are placed, despite the biological evidence that such categorization at a particularly early age should have little to do with later achievement.

This persistence of tracked groups, whereby the initially 'slow' students tend to remain slow despite the fact that social or physical development may speed up, is perhaps among the best evidence that tracking students is not simply a convenient pedagogic device but results in self-fulfilling and often damaging value judgments. There is absolutely no reason to assume that a child who learns to read late will be a poor reader, just as there is no reason to assume that a child who learns to speak later than another will be a poor talker. There are

many parents who have discovered that a late start in talking has not prevented their child from becoming a voluble and constant chatterer later on! If all of those students who start slowly in certain school skills stay slow, it suggests that the result is due more to their school classification than because of biological necessity.

A second related point is the common knowledge about the correlation between social behavior and school track. By and large, the highest tracks in a school--that is, the fastest children--are usually the best behaved, while the slowest are the poorest behaved. Again, there is no reason to believe that the slow development of mental processes alone has any relationship to behavior. No one assumes that children who are slow to learn to walk or talk, or who grow slowly, present more serious behavior problems than those who do this rapidly. If children who happen to be slower in development of reading skills than others, for example, end up behaving badly,the reason may well lie in the way they are treated because of this developmental trait.

A respect for children and a desire to see them develop to their fullest potential also requires cooperation and mutual interaction in learning, rather than competition and isolation. Educational environments should maximize opportunities for children to become conformable with the world, to face it, to structure and order it. To the extent that schools rigidly classify subject and process for children, they deny children the experience they need in order to organize the world. If we accept that experience is necessary in order to understand the world, then schools must endeavor to help children to understand the connections between things by enabling them to make these connections. For example, there is a relationship between spelling and writing and reading, but it is a lot harder to understand what it is if these 'subjects' are always taught in isolation or in a particular order. There are certainly connections and tremendous overlap between art and science and crafts, connections which are variable and of different significance to particular people. But the only way for any of us to be able to make those connections for ourselves is to have the opportunity to make cross-links through our work.

In a similar manner, it is only possible to learn about the social world by participating in it. School--if it is an educational institution--must support social interaction. This means fostering cooperation, sharing, assistance and all forms of social relations between people: children with children, children with adults, and adults with adults in ways which allow the individuals concerned to get to know each other better and to learn cooperatively from each other. Most competitive situations have just the opposite effect: they draw people into themselves, encourage them to become suspicious of others, to keep things to themselves--in short, they isolate people from each other. This isolation discourages children from learning and growing.

Competitive situations also are inconsistent with appreciation for individual differences in growth and style. Some children read faster than others but not therefore necessarily better. (The same is true of adults: reading speed has very little correlation with intellectual training or ability, or, for that matter, with retention of what is read.) Some children are good at spelling out loud, others are poor at it. Most important, some children can do any number of these things well at some times and not at others. The point is that stress on isolated measurement of particular skills does not really enrich our knowledge of children's growth. Instead it tends to make us stereotype children in respect to a few properties and to forget to ask what they are like as people and what other strengths, weaknesses, and interests they may have.

2

General Position Statement on Evaluation

1. Children go through stages of physical, mental, emotional and social growth, and it is important to know where the children are, at a given point, and what one may expect next. In many instances, these stages can be expressed in a quantitative manner. There is no point in saying that Susan is short or that she is just so high; one can readily report that she is 42-inches tall. Likewise, abilities to read or compute can be described with some precision. However, it is always important to recognize just what those measures mean. Being 42-inches high at age six is a fact, but one that has little relation to worth or general ability, or even to how high Susan can jump or whether she can run fast. Likewise, a sight vocabulary score is just that and nothing more.

Perhaps most important to stress in discussion of quantitative measurement is that we are interested in these measurements because of what they tell us about that child, not because they permit easy comparisons. This point is at the heart of all discussions about evaluation and measurement. Any statement about a child's achievement and level tells something about *that* child, and is significant in a description of that child. So first and most obvious, evaluation results should not be formulated in terms of averages, but in terms of individual results. Or, to put the same statement in somewhat more technical language, what is interesting in any group measurement is not the mean but the variance.

It is also important to remember that in the assessment of any evaluation effort, what has actually been measured must be clearly stated. A measurement of height or weight is direct and we know what it means; many educational evaluations are not. For example, students who can read quite well and comprehend what they read, may receive a low score on a reading test if they do badly on some technical sections, such as blending, syllabification, auditory discrimination (Allen, 1974).

2. Evaluation practices must respect the setting in which the educational effort takes place. That is, it is necessary to adapt the evaluation to the program rather than *vice versa*. When the educational endeavor is one which advocates learning through interaction with the world and through social interaction, it becomes particularly important that the measurement of children's growth also involves the 'stuff' of the world and permits social,

cooperative interaction. The whole process of evaluation must also take into account the effect of the testing itself on the school setting and on the whole program. This is important both in the measurement of children's progress and in the evaluation of programs.

3. All evaluation efforts should recognize the distinction between saying and doing, between verbal knowledge and ability to use information. If I want to find a good mechanic for my car, I usually don't ask the mechanic to describe the internal combustion engine. If I want an electrician, I don't ask for a definition of electricity; I want to find out about the work of these people. Similarly, evaluation of children's work should take into account the doing of that work, not merely descriptions of it.

4. The value of any evaluation is in direct proportion to its usefulness, to how much it can help a child's education. If there is any measure of what a child can and cannot do, then this should be in a form and at a time when it can directly assist the people who are working with that child.

To summarize, any evaluation of children's performance, whether quantitative or qualitative, should stress the individual results rather than make comparisons, express these results in a manner that is useful to the people involved, relate it to the particular educational setting, and recognize that children are complex beings with a wide set of attributes and influences.

3

Present Status of Educational Evaluation

Evaluation is making judgments about a process; educational evaluation involves making judgments about a social, public activity. Examples of evaluation questions are: Is this school adequate? Is this a good teacher (principal, administrator)? Is child "x" making reasonable progress? Should we use curriculum "a" or "b"? The way to arrive at these decisions is to use the best and the most information possible. One major source of information is some kind of measurement; or, to put it the other way around, a particularly useful way of gathering data necessary for making judgments is to make measurements, to collect data, and to present it in some orderly form. But obviously this is only one component of responsible decision-making.

The passage of the Elementary and Secondary Education Act in 1965 marked the first major direct introduction of federal funds into the public school systems. Much of this money was allocated for programs directed towards poor and minority children. The advent of this intense effort of federal money spent on the schools brought with it a sudden cry for 'accountability'--for finding out whether the money spent was doing any good. Although it is easy to understand why questions should be raised about the expenditure of federal money for education, as in other areas, it is worth noting that such questions were first seriously raised only when money was beginning to be spent in poor districts and to alleviate the educational shortcomings of poor and oppressed students.

Also the nature of the questions was of a very interesting kind. The major issue was not whether the money was spent as the law required, that is, specifically for reading improvement or for the arts or for bilingual education, but whether the money was actually solving problems that existed in the schools. In other words, the stress on *evaluating* programs focused on the results that might arise, not on the way the money was expended. A comparable situation would be if the massive highway fund had been evaluated in terms of whether it solved the transportation problems of the United States, rather than whether the money had actually been spent on highway construction, labor, cement, steel, etc.

In response to the outcry for evaluation, the educational community brought to bear its best and brightest

minds concerned with evaluation. The field became more visible, and considerable amounts of writing and prescription followed. In 1967, the American Education Research Association (AERA) began to publish a series of pamphlets on the subject of evaluation. Several of the articles in the first issue discussed 'professional' evaluation and urged strongly that evaluation studies not be left in the hands of amateurs but entrusted to professionals.

It is interesting to think about who the counterparts to 'professional' evaluators would be in other fields of human endeavor. When the stress is put on the measurement part of the work, as it often is in the literature of evaluation, then it is tempting to think of evaluators as the analysts of the field. By this definition, they are the people who do work comparable to chemists who analyze compounds for their elemental components: the amount of carbon, hydrogen, and nitrogen in a compound. But the analytical chemists by themselves do not make judgments. They simply follow a procedure established by practicing chemists and report results from the procedures; they certainly do not, or should not, make value judgments. A chemist would be surprised and annoyed if she received a report from an analyst which said, "the compound you sent me contained 67 percent carbon, 8 percent hydrogen, 19 percent nitrogen and isn't worth reporting in the literature!"

In education, professional evaluators have a role which is much more like management consultants, consumer advocates, or any of that range of people who try to look at some social activity critically and then make judgments about it. If there is one thing we know about this whole kind of activity, it is that although good judgments require careful collection of data and measurement as a necessary activity, this is hardly sufficient. In fact, when the preoccupation with details and data gets too great, then the most important issues can be forgotten. In the *Best and the Brightest*, a book about the Vietnam war, David Halberstam points out the limits of great professionals hard at work on a social problem. In describing the decision-makers in the Pentagon during the war, Halberstam reports that while there was much analysis, great gathering of data, body counts, and constant reports from Vietnam, some simple general truths tended to be forgotten, and crucial questions remained unexamined. Consequently the data proved over and over that we were winning the war despite the contrary evidence.

A second issue concerning professional evaluators has to do with their background. If there is such a thing as professional evaluation, then the members of this profession must have been trained somewhere, or must at least have some identity as a profession and some views and ideas they share as members of this profession. By and large, people who call themselves professional evaluators in education have been trained in social science research in universities. Most of them come from educational psychology departments or similar departments with other names. Their

reference point is the American experimental psychological tradition, especially as practiced in the field of education.

Educational psychology, like every field of science, has its own style of operation and its own way of defining experiments, goals, and approaches to problems, even its own style of defining what a problem is. What appears to be a reasonable approach in one field, however, is just not acceptable in another. American experimental psychology, with its strong behavioristic strain, has developed a particular scientific tradition, with its own norms, methods, and goals. But this is simply not appropriate to the entire range of human activity. Schools are social institutions carrying on a complex social and cultural activity, they are not experimental laboratories in which controlled conditions can be established and isolated events studied relatively separate from their surroundings.

For some years I taught chemistry and biochemistry in a large urban university. I taught undergraduate courses in organic chemistry and supervised graduate students doing biochemical research with enzymes. We published papers in respectable journals, received federal financial support and the students who worked with me successfully competed for scholarships and professional recognition. Yet, a few of my colleagues in the department consistently told me that I wasn't doing 'real' research, that biochemistry wasn't a 'real' science, and that my students weren't getting 'good' or sufficient training. The only way to satisfy these colleagues (who were in the more physical end of chemistry) would have been for us to give up our particular interest and to adopt theirs, along with the techniques, the particular mathematical tools, and the general styles of approach which appeared to them as the only appropriate ones. Of course, my critical colleagues in physical chemistry were being told by some physicists that all chemistry was just a minor, imperfect, and not very important branch of physics, and that only the physicists were doing 'real' science. As Kurt Vonnegut would say, "so it goes."

The experimental psychological tradition is at some point in this spread of the range of science, with its own particular models and techniques. Whether academic research in experimental psychology is the best model for evaluation studies is a question. Before going into it, however, we have to ask whether *any* traditional academic research style is an appropriate model.

The moral dilemmas that enter into 'pure' scientific research, such as basic physics, or chemistry, were made painfully obvious to society by the events surrounding the Second World War. Recently researchers in biology have argued that certain experiments simply should not be carried out until the possible social consequences are evaluated. These moral and social questions are constantly troublesome in all uses of social science. Unfortunately, traditional descriptions of scientific method are based on views that fail to take these factors into ac-

count. For centuries, scientists have developed a style of 'objectivity' and a set of methodologies that have ignored the social implications of research.

This discussion assumes that academic research is carried out by a specific set of rules and that evaluation work also follows these rules. This is a fairly standard textbook view of science and of activities of all sorts: that there are some right ways of doing it and some wrong ways, and that people who carry out the work do it correctly, or else are frauds or failures. Of course the world of actual practice doesn't work that way. There are some 'proper procedures', some 'correct' ways of carrying out anything, whether it is repairing cars, running a factory, or doing research. But these correct ways change with time, and more important, anyone who does work well knows there are times when you simply throw the rules out the window and do whatever you have to do to get the job done.

Moreover, particularly significant measurements sometime require new instruments. Part of Galileo's problem in convincing people of his evidence for the organization of the solar system was that he was using a whole new measurement technique: he was looking at the moons of Jupiter through a telescope. Was that a legitimate measurement device? People had to decide whether it was or not. Similarly, various indirect ways of looking at nature--measuring electrical charge, spectroscopy, radio waves--had to be accepted as legitimate measurement devices. In many cases, the advent of a new bit of science or technology required that the new way of measuring also had to be invented and then accepted as part of the proper instrumentation.

4

Some Characteristics of Evaluation Paradigms

THE DOMINANT MODEL

It is worthwhile to look at the general nature of the research design methodology that dominates measurement in the field of educational evaluation in order to decide just how relevant (or inappropriate) it might be.* The general model comes from that branch of psychology that attempts to model itself on research which proved particularly successful in 18th century physical science, and was then applied in the 19th century to more practical problems, and to areas which needed a little manipulation in order to meet the same criteria for research. Each field of science has particular methodological issues which are difficult, and others which are relatively simple depending on the nature of the subject matter. In observational astronomy, for example, it is relatively simple to carry out and standardize repeated direct observations. The phenomena in the sky are uniquely there. They repeat themselves, and all one has to do is sensibly and patiently observe. Also, the phenomena are accessible everywhere on the earth, relatively stable for centuries, and similar over large areas of the earth, so that checking observations from one point in space or time to another point in space or time is quite easy. On the other hand, some kinds of experimental work in astronomy are virtually impossible. Bits of the heavens cannot be isolated in order to take them into the laboratory and change the conditions to see what happens.

One of the triumphs of late 19th and early 20th century science is the devising of mathematical techniques and experimental tools for work in fields where the actual number of individual bits of experimental materials is not vast as it is in chemistry, or regular and beyond reach as in astronomy, but relatively small in quantity and able to be manipulated. Some of the best work in this area was in the field of agricultural research, and the classic studies in research design now widely applied in the social sciences still refer to these methods. A standard reference is the work of Fisher (1935), who summarized the methodology recommended in the 1930s.

In this approach, an experiment is defined in terms of taking two populations, selected at random, doing something to one of them, using the other as a control, and comparing the two before and after the treatment. By im-

*See also Michael Patton's *Alternative Evaluation Research Paradigms* in this series.

plication, this approach becomes *the* method of doing research, the only method of arriving at new or certain knowledge.

Because of the nature of the populations available in social science research, two problems--that of selection of experimental and control groups, and the relationship between experimental treatment and results--assume an enormous significance. I want to discuss this research paradigm in terms of its relation to educational evaluation.

Applicability.

In this research design, considerable energy is expended devising ways to arrive at a random sample to make sure that the population studied is some average general one, not the result of some prejudice or odd local factor that might influence the result. A good deal of agricultural research exemplifies this point. If you want to find out the effect of a particular fertilizer on corn crops, you must be sure that you don't confuse the effects of fertilizer with the effects of rain or weeding. Also you want to know just how great the effect is.

The influence of this particular way of doing research in educational work is indicated by the writing in the field. Many theoreticians and methodologists describe it as if it were the *only* possible way of doing research. In "Experimental and Quasi-experimental Design," a highly respected outline of research methodologies (Campbell and Stanley, 1963), the authors recognize that there are many cases where the 'standard' of Fisher-type experiments cannot be met, but they make it clear that such situations are, at best, 'quasi-experimental'.

The basic problem is not with Fisher's or Wendell and Stanley's definition of an experiment. They are at liberty to define this as they wish. What is frightening and limiting is the further suggestion that experiments defined in this way are the only way to acquire knowledge, or that, no matter what the situation, every effort should be made to structure situations so that experiments of this kind are undertaken.*

An example from the evaluation literature illustrates the contempt which persons concerned with educational evaluation have for whole fields of scientific endeavor. In the AERA monograph (1967) mentioned earlier, Michael Scriven writes:

> We might for example be interested in the proportion of the class period during which the teacher talks, the amount of time that the students spend in homework for a class, the proportion of the dialogue devoted to explaining, defining, opinion, etc. (Milton Meux and B.O. Smith, 1961). The great problems about work like this are to show that it is worth doing, in *any* sense. *Some* pure research is idle research. The Smith and Meux

*One of my favorite stories about research work concerns a bright young biologist who wanted to repeat some experiments carried out by an established researcher on a particular strain of microorganism that the older worker had isolated and cultured. The young man wrote to his senior colleague asking for a sample, and was turned down. He continued to write, constantly renewing his request, although all his letters were received with negative responses. When a colleague asked the young biologist why he kept repeating his request when he should know that the answer was going to be "no," he replied, "I know he will refuse me, but his letters are written near his lab, and everytime I get one, I cut it up into little pieces and see what I can grow from it on agar plates. Sooner or later, I'll get my organism." This imaginative and outrageous bit of methodological strategy just doesn't fit into the models of neat experimental design.

> work is specifically mentioned because *it* is clearly original and offers promise in a large number of directions. Skinner's attack on controlled studies and his emphasis on process research are more than offset by his social-welfare orientation which ensures that the process work is aimed at valuable improvements in control of learning. It is difficult to avoid the conclusion, however, that most process research of this kind in education, as in psychotherapy (though apparently not in medicine), is fruitful at neither the theoretical nor the applied level. (p. 50)

The implication here is that 'process research', that is, field studies based on observational methods, is not even worth doing unless it is offset by a particular social-welfare orientation.

This is a rather harsh judgment on a large number of scientific fields, which might have particular relevance to evaluation. Anthropologists, archeologists, ethologists, a whole range of social scientists do not do 'controlled experiments'. The basis of their work is informed observation, and it is remarkably fruitful. Jane Goodall (1971) watched a small number of individual chimpanzees over a period of years, and in the course of her observations discovered the apes using, and even making, tools. She could not possibly have developed a control group experiment; in fact, she would have had no reason to set up such an experiment, even if it were possible, because tool-making was not part of the expected behavior of chimpanzees. Because she is in a field that accepted open observation without a specific predetermined behavior being measured, she could make her scientific discoveries.

Another difficulty in trying to apply the agricultural, experimental research method to evaluation arises from the fact that evaluation is *not* a laboratory research activity. It is performed in the field; that is, in natural settings--in schools with live children and adults. This makes the whole problem of randomization extremely difficult, because the total environment cannot be manipulated for either experimental or control groups.

What Data is Generated.

The fact that this experimental methodology is hard to apply is not, of course, sufficient reason to question it. But one can ask whether the information it yields is worth all the trouble. The kind of results that can be obtained by applying experimental designs as described by Campbell and Stanley give only a small part of the types of data that are needed to make educational judgments. This method focuses on comparison of averages, means, total sample gains, and generally on trends which apply to the group as a whole. It is not designed to focus on individual members of the experimental sample.

For example, in determining the effects of fertilizer on a crop you look not at individual plants, but large fields and the weight of the resultant crop. It may be the case that a fertilizer produces amazing results by stimulating 90 percent of the individual plants, although it kills the remaining 10 percent. This can still make a fertilizer highly desirable. But imagine a similar situation in education!

In some cases, it is useful to obtain the kind of data that is generated by applying the experimental psychology model of research. The Plowden Report (1967) provides an example of a situation where data gathered by the statistical research methods of 'experimental design' was exactly what was needed--within a context. One question the Plowden committee asked was simply: what is the general level of reading attainment of English children as compared to the same group several years earlier? By giving standardized reading tests to a fairly small sample, the committee was able to determine that reading levels for the whole population had increased over the time period measured. That is a case where the question asked was best answered by this kind of impersonal, averaging procedure. The desired information was general and impersonal, and it only required sampling a small fraction of the entire school population. Still the major work of the Plowden Committee, to analyze the status of primary education in England and make recommendations for the future, depended on interviews, observation, and concrete examples.

View of Causality

Another problem with the style of the dominant research paradigm is that it is based on a rather naive and simplistic notion of the nature of causal relationships in social situations. The basic premise, derived partly from the behavioristic outlook of many research predecessors and partly from the kind of methodology that is advocated, is that there will be fairly direct and immediate results from particular actions. Introduce program "a," teaching method "b," or organization of classroom "c," and it will be possible to see the effects fairly directly and separately from other events that may happen. It is easy enough to see how this view can be applied in the study which served as a model for this type of research. Plants are relatively simple biological species, they don't particularly interact with each other, and they have relatively few degrees of structural freedom. Therefore, they respond simply and directly to specific changes in conditions. If you water them more, they grow more. If you fertilize them, they grow and produce more, etc. But people just don't respond that simply to stimuli, especially not in open, natural social settings.

The kind of methodology classified as 'experimental' by Campbell and Stanley is based on the assumption that the effects of actions can be isolated and measured fairly directly, and that what is going to hap-

pen can be predicted with enough precision to look for that result and directly relate it to isolatable situations, programs, and activities that you wish to evaluate. It is possible to try to arrange human studies to approximate the simple conditions of plant life, but besides the terrible moral issues which then come into play (see below), to attempt this is to destroy the real world situation of ongoing school activities. It is this ongoing work which is the proper subject of evaluation studies.

Moral Issues

A crucial issue of any research strategy, especially any which involves living things, is the moral problem involved. What does it mean to do any sort of research that includes humans? There appears to be a common misconception that as long as research follows proper methodology, such questions are resolved or at least minimized. Certainly some of the writing about proper methodology appears to ignore the implications of these positions. For example, in developing an argument to show that it is possible to do comparative studies even in cases where absolute results cannot be obtained, Scriven* states:

> The analogy in the medical field is not with drug studies, where we are fortunate enough to be able to achieve double-blind conditions, but with psychotherapy studies where the therapist is obviously endowed with enthusiasm for his treatment, and the patient cannot be kept in ignorance of whether he is getting some kind of treatment. If Cronbach's reasoning is correct, it would not be possible to design an adequate psychotherapy outcome study. But it *is* possible to design such a study, and the way to do it--as far as this point goes--is to use more than one comparison group. If we use only one control group, we cannot tell whether it's the enthusiasm or the experimental technique that explains a difference. But if we use several experimental groups, we can estimate the size of the enthusiasm effect. We make comparisons between a number of therapy groups, in each of which the therapist is enthusiastic, but in each of which the method of therapy is radically different. As far as possible, one should employ forms of therapy in which directly incompatible procedures are adopted, and as far as possible match the patients allocated to each type (close matching is not important). There are a number of therapies on the market which meet the first condition in several dimensions, and it is easy enough to develop pseudotherapies which would be promising enough to be enthusiasm-generating for some practitioners

*Repeated reference has been made to this particular article because it is one of the most influential documents that has appeared, and continues to be included in anthologies.

(e.g., newly graduated internists inducted into the experimental program for a short period). The method of differences plus the method of concomitant variations (analysis of covariance) will then assist us in drawing conclusions about whether enthusiasm is the (or a) major factor in therapeutic success, even though double-blind conditions are unobtainable. (p. 68)

The implications of having eager, inexperienced young internists practice pseudo-therapies on innocent, but troubled, patients involves serious moral questions. Similar suggestions have been made and should be resisted in education. There are cases where children are simply denied benefits for the sake of setting up a control group, or where for the sake of completing the experimental activities children and parents are not fully informed of a program. Adherence to a 'good' research design, that is, one that is methodologically sound, does not even begin to address any of the moral questions that come up in a particular research activity.

ALTERNATIVE STRATEGIES OF MEASUREMENT

There are other scientific research methodologies that provide approaches to the collection of data, which are particularly appropriate for many evaluation studies. Increasingly in the last few decades research on schools has used the approach of the anthropologist, the questioner of culture, to examine what happens and to describe, tabulate, appraise, and finally, judge or evaluate education (Kimball, 1972). A good part of this work was inspired by Jules Henry (1963) and his general anthropological approach to looking at institutions. Since then, a number of people have applied similar methods. Philip Jackson's book, *Life in Classrooms* (1968), is a proper, scientific research study, but its research methods come from a different field than behavioristic psychological research.

The power of the anthropological approach can be estimated from the impact that this type of study has had on American education in recent years. Serious discussion of the educational scene has been generated by the descriptive indictments of the schools contained in books that range broadly from popular and impressionistic works like those of Herndon (1965), Kozol (1967), and Kohl (1967), through the personally analytic like Holt's *How Children Fail* (1964), to the more scholarly, such as Jackson's book and Ray Rist's *The Urban School* (1973). All these efforts have two things in common. They describe the schools from an anthropological-sociological perspective, and they paint an intensely gloomy picture of school life. No experimental study, in the Fisher sense, could provide this information.

If educators and the public want to evaluate a

school or some aspect of school life or individual children's growth and learning, they have to apply the tools that will give them the most and the best information. This requires surveying the entire field of social science to pick out what is appropriate. For any major task involving evaluation of new activities, it probably also means inventing new ways of getting the information.

5

Classification of Evaluation in Education

The term evaluation is applicable to a range of activities which require judgments to be made. For the purposes of our discussion, it is possible to organize and discuss these activities under three headings. Each has its own problems and its own techniques, but they also involve similar issues.

First is the general area of evaluating the growth and development of individual children. This is, of course, what American schools are about, at least what they are supposed to be about formally--fostering the growth and development (the learning) of individual children. This is also the area in which the experience of half a century of tests and measurement in experimental psychology has been most directly applied.

A second level of evaluation concerns judgments about the performance of various other people in the school system. Evaluators ask questions about how well teachers are doing, how well principals are acting, who should be hired as a school superintendent, etc. Judgments of this sort are related to, but not identical with, knowledge about children's growth and development. The kind of job that is being done by the teaching profession as a whole, or the kind of service we are getting from school administrators as a group, is and should be reflected in the reports we get concerning children's growth into healthy and competent adults. On an individual level, this generalization breaks down--there are simply too many factors involved. It might be possible to draw conclusions about the type of health service in the United States from surveys of the general state of health of the population. It would be much more difficult to make judgments about the competence of each individual doctor on the basis of the average or general health of her patients.

Because the measurement of individual children's results on standardized tests is sometimes the only concrete evidence that is gathered about the way a teacher is doing her job, there is a tendency to judge her on the basis of those results alone. Making such judgments on insufficient data, and without careful thought is a rather dangerous practice. The recent fad for performance contracting, in which school personnel and programs were judged by the results as measured by pupil performance, is an example of this practice. The unhappy events that resulted (Report to Congress, 1974) parallel what occurred in England late

in the 19th century when teachers were paid according to the examination results of their pupils. In both cases the system encouraged a surge of improprieties: teachers and administrators saw to it that a minimum level of 'performance' was guaranteed, no matter how it was obtained.

The third general area of judgments is the evaluation of programs. At the minimum level, this involves judgments about a new curriculum, some educational innovation, or other specific program brought into a school, such as Title I, Title III, NSF-sponsored science and math curricula. Much of the recent stress on the importance and necessity for evaluation is a direct result of the federal expenditures in education which were directed at programs of this sort. Here the situation is similar to the one that prevails in judgments about school personnel. Decisions are desired concerning total programs; the methodology that is available is about individual or average pupil achievement on standardized tests. Attempts to connect the two by some simple causal relationship may not be applicable. There are a number of such discrepancies between federally sponsored programs and standard evaluations available. Many of these programs have goals that are quite different or much broader than those encompassed by standardized tests of achievement developed within the psychometric models available. Yet, there is often an attempt to assess the programs in terms of these tests. To refer again to our medical analogy, it is as if the success of a variety of community health programs were all measured on the basis of a set of standard measurements on individuals concerning their general health: blood pressure, weight, number of operations, etc. This might be an appropriate, although not a sufficient, evaluation for community health programs related to sanitation or diet, but it would certainly not be the most useful information for judging a program that had as its goal the development of mental health centers or family planning, or addressed other broadly conceived health issues.

Program evaluation also concerns integral parts of school organization within a single policy unit. Thus, for example, a new curriculum or program may be tried out in several classrooms, or within one school, or in part of a district. Again, evaluation judgments have to be made, and again, the results for individual children are part of, but not all, the information needed to make a judgment. A complex of social, political, and economic factors need to be taken into account. An education program that, for example, increased reading scores for children at the expense of their physical health would be highly unlikely to be acceptable, no matter what the standardized test results showed. Similarly, any program or innovation that disrupted the health of the school system or the functioning of the community would have serious problems; the judgments about it would reflect this, regardless of what the intellectual growth and development of individual children might be.

Questions about general educational policy represent the most complex level of programmatic evaluation. These involve decisions about curriculum or educational goals and how they are determined locally or nationally. It should be most obvious in this instance that results obtained on standardized tests of children's academic achievement are only a fraction of the information needed for sound judgments. A major function of every school system is the socialization of the young into the society. Only a fraction of that socialization is concerned with academic skills; standardized tests are not complete measures of academic achievement. Thus, it is simply not possible to make judgments about the important functions of the schools on the basis of individual pupil performances in achievement tests, apart from their social and cultural context. A similar critical appraisal of the relationship between standardized testing and program-related assessments is found in a recent publication jointly sponsored by the National Institute of Education and the National Council of Teachers of English (Venezky, 1974).

The reliance on achievement testing of children to evaluate a wide range of educational practice is so remarkable that one has to wonder why sensible people would even advocate it. Why should a teacher, who has responsibility for many things besides the academic achievements of children, be judged only by that achievement independent of her working conditions, support, local problems, school system goals, social pressure, and her ability to inspire or teach or guide or socialize children as is proper for that community? Why should a school system,which is charged with keeping children out of trouble, satisfying a community expectation, providing recruits for the labor market, training consumers, and a host of other tasks, be judged only by the achievement on standardized tests of the children in the system?

It *is* possible to make some judgments about the nature of the society and the nature of the role school systems play from comparative achievement data between parts of the population. Perhaps the most striking value of the achievement tests,which are so widely used by the schools,is that they give solid, 'objective' proof that the schools support the racism and discrimination that exists in American society. The one standard measure that our society uses in judging our children and our education system shows conclusively that we have created a system that hurts a large fraction of the population, much of it black, and most of it poor. The fact that broad categories of students--urban blacks or poor whites--have systematic non-normal distribution of test results, on tests that have been designed to provide normal distributions, clearly illustrates that our society is treating groups of children differently and then determining their future on the basis of this treatment.

6

What Do We Know About Children, and Why?

The appropriate basis for developing any program to evaluate children's growth and development is to decide what it is, in fact, a particular audience wants to and needs to know. Also central to any evaluation decision is the question: for whom is the information being gathered? For centuries there was a struggle to free science from what appeared to be irrelevant, and often stifling, political and social considerations. Unfortunately, this struggle, along with the general scientism of the late 18th and 19th centuries, led to the belief that whatever is studied in a 'scientific' manner is divorced from any social or political considerations whatsoever. It even made the question--who wants to know and why?--an irrelevant one. During a period when science was the plaything of educated gentlemen, this disregard for social implications of the uses of science was perhaps possible. But recent history has made us aware of the social and political uses to which various forms of scientific enterprise have been directed. We need to be concerned with such matters as who is interested in poison gas, or who wants to know about psychological methods of persuasion, or why a government agency is collecting data about citizens.

Although the motives and reasons for obtaining educational evaluation data are usually not as sinister as in some of the examples I have cited, we also have to ask who wants particular information about children and why it is requested. The purposes of an evaluation effort and the audience for whom it is intended is often a guide to what is and is not appropriate information.

THE NEEDS OF TEACHERS

1. One area of concern for teachers is whether children are learning direct, specific skills: sounds of letters, mathematical operations, rules of kickball, or how to look up the spelling of a word in a dictionary. This kind of information, in many cases, can be easily determined by using standardized tests. It is quite possible to give a child a test that will determine if she can read the word "ball" or give the correct answer to the question 3 + 3 = ? But in most cases, a teacher can also obtain the same information quite easily in other ways in the course of day-to-day contact with the children. Any

teacher who has children read to her will get a reasonably good idea what words a child knows. Any teacher who plays a board game with children that uses dice, for example, will find out about a child's ability to add numbers up to 6 + 6.

In discussing the use of reading tests, Venezky states:

> The number of different instructional groups into which students are placed is generally small, and the differences in predictive ability of even the most extensive formal tests over informal teacher judgment have never been shown to be large. (p. 7)

2. Information is needed about children's more fundamental growth through stages of development. Increasing vocabulary or learning more 'number facts' does not constitute advances in the kind of thinking the child can engage in. On the whole, it is not possible, using simple questions with multiple choice answers, or true and false, or fill in the blank, to obtain information about how an answer was arrived at, the reasoning process that was used to arrive at an answer, or the levels of complexity that are involved. For example, it is fairly easy to devise a method to determine how long a number anyone can remember. You simply ask the person to repeat a number back to you, starting with a one digit number, then a two digit number, and so on. But if you want to determine a person's reasoning process, the task becomes much harder, if it is possible at all. As problems become more complex and more interesting, the ways to attack them also increase in complexity and in number so that no matter how carefully you structure the problem in parts, there is simply no way--looking only at the answers--to find out how a person arrived at the various responses to complex questions. Any experienced test taker knows the strategy which argues that a particular answer must be correct (or incorrect) because it is the kind of answer that would be expected by that particular test or tester, or because the answers on this type of test are bound to be whole numbers, or because there wouldn't be two similar answers, or--a strategy that a friend of mine swore she used with great success--"in all multiple choice exams, if one answer is significantly longer than the others, it is always the right answer, because no one would bother to make up a long wrong answer."

3. Horizontal Growth: Related to the question of developmental growth--how children think, how complexly they can approach a problem--is the issue of horizontal growth discussed earlier. How rich is the experiential base and how rich is the thinking on any one level? This is such a subtle and under-explored area of development that there are obviously no simple ways to get at questions about it.

4. Another area of concern for teachers is how well children can use the skills they have. There is no simple

correlation between mastery of vocabulary and syntax, and actual reading done, or even reading with comprehension. This question of use of skills requires both the skill itself and some sort of context in which to use it. For reading, it requires actual reading; for math, quantitative manipulations; for art, the production of something expressive; for crafts, construction of objects; for sports, participation in the activity. It is not clear that situations that are specifically designed to test the use of a skill outside the context of actually doing something have much relationship to that skill. It is certainly not enough to look at the results of tests to find out whether children actually use certain skills.

5. Finally, there is the question of 'learning to learn', or learning problem-solving, heuristics, or any of a number of terms that have been applied. Increasingly, educators are becoming aware that education should strive to develop in people the ability to take care of themselves, to undertake their own continuing growth and development, to deal effectively with situations not previously encountered.

Information that is useful to teachers is direct, immediate, and specific about the children in this year's class. From the viewpoint of an elementary school teacher, the disadvantages of national standardized tests far outweigh their advantages. Of the five areas discussed above, only the first is covered in these testing programs. But even this information is given in normative terms, comparing a child to some national sample, rather than in individual terms, helpful in working with that particular child. Also, it takes an enormous amount of time to get back the results. At the same time, the tests disrupt the educational work in the classroom, demoralize and disturb the children, and disrupt the helping relationships established among the children and between the children and the teacher.

THE NEEDS OF SCHOOL ADMINISTRATORS

The situation changes when we look at the kind of information that school administrators find useful. It might be hoped that school principals, as part of the task of supporting teachers and being concerned about the education of children, would be interested in the same sorts of things that teachers need to know about children. Actually, most American school principals are not head teachers, but organizational administrators; they are concerned with staffing, busing, and discipline. They don't have the time, training, or, for the most part, the inclination to teach and to be involved in the growth of individual children. Since principals, and the rest of the administrative ladder in a school system, see themselves as supervisors of a system rather than guides for individual children, they need system-style data: short, concise, and easily compared. Also they are concerned

with trends: How does performance or learning, or any other measure, compare year-to-year? How does it relate to expenditure or to changes in practice, etc.?

Concern for trends and comparative data is necessary within any school system. It is necessary to know as precisely as possible how a particular practice affects results; that is what evaluation is all about. Unfortunately, for the reasons indicated, the information obtained from standardized testing is simply inadequate for many of the decisions for which it is used, or at least for which its use is claimed.

THE NEEDS OF PARENTS

Parents are also interested in information about the school. On the one hand, parents want information about how their own children are progressing in school, what they do, how they behave away from the home setting. On the other hand, as community members and tax payers, they want comparative information and information on trends in the schools-at-large. In situations where they have become involved, parent concerns usually go far beyond the limits of what is provided by standardized tests.

Parents want to know what their children's chances of success in life might be. It is often argued that one of the reasons school systems need standardized testing is to give parents this information. If educators didn't have these scores, parents would not have much idea of what their children's education was worth, what it could do for their children in the long run. Pressure from parents is often said to be the reason reading scores are stressed because parents believe that reading scores are related to future success, to getting into college, etc. It is ironic that school personnel should point to parents as the force that supports the tests, because it was the school administrators and academic experimenters who originally sold the tests to the parents. The great trend towards quantified statements of school performance came not from parent and community groups but from the scientism of the academy in the early years of this century (Cremin, 1961).

The problem with the belief that individual high test scores lead to success in society is that it introduces the lottery concept into education. The possibility of high test scores is held out to low-income parents as a way to provide a great future for their children when, in fact, it would take a very high score indeed to change significantly the life chances of poor children. The relation of school to college admission, jobs and income is a complex one closely related to the prejudices and discriminations in our society (Berg, 1970). It is true that an unusually 'high-achieving' child from deprived circumstances--that is, a child who does very well on the standardized tests--can break out of the bounds of the economic and social class in which she lives, and ac-

tually change her status. But the odds against this are enormous. This kind of case--and there are some all the time--has the same effect on redistribution of classes in society that the lottery has on redistribution of income. The lottery in Massachusetts, for example, provides about a 13,000,000-to-one chance against winning $1 million. That means that after 13-million tickets are sold, one person may significantly change her economic status.

There are just enough winners of smaller amounts so that many people can support the illusion that they too may be a winner, that they too can change their status. But, of course, the actual number of people who do win something is so small as to be insignificant for any change in class alignment. Exactly the same reasoning holds for the concept that good reading scores will help populations break out of poverty or oppression. The actual number of children who can change their status as a result of school success is trivial compared to the total population that is condemned to poor jobs and continuing poverty.

An analysis of the kind of information different groups need leads to the conclusion that the present system of reporting children's standardized achievement scores, at best, only assists school administrators, and only in one of their functions: that of acquiring data for long-term planning and assessment. Even in this area, the results available are one small component of the information that is needed for intelligent decision making.

7

The Measurement of Child Achievement

Because the measurement of childrens' achievement represents the most widespread standard practice in the schools, because it is often the basis for many other judgments, and because it is at the primary level of the evaluation hierarchy, I would like to examine it in more detail.

STANDARDIZED ACHIEVEMENT TESTS

One of the most remarkable features of the present state of the measurement of child growth and development in the schools is that while all thoughtful educators agree that the available tests are terrible, almost everyone continues to use them. Meeting in Washington in 1972, educational sponsors of Follow Through programs agreed unanimously that the available tests were inadequate to measure what was happening to the children in Follow Through classrooms. Strong criticism was expressed not only by the sponsors who advocated more "open" programs with emphasis on varied learning style, affective development, and social concerns, but also by the sponsors who advocated more traditional programs. Many of the sponsors were highly critical of Stanford Research Institute, the organization hired by USOE to conduct the official overall evaluation, for not developing more imaginative and useful measures of child growth and development. Yet, for a number of reasons, including social and political ones, many of these sponsors actually used identical tests in their own evaluations!

The usual generalized argument given in support of continuing to use recognizably inadequate tests is that there is nothing better available. This argument is a sound one if an activity or a process is simply not as good as it could be--that it is inadequate. But criticism of standardized achievement tests goes much further that that: the tests are not only not good enough, they are harmful and destructive to a number of school programs; they are especially harmful to children.

A number of specific points, some of which I've touched on in passing, can be made in this regard:

1. The present tests are discriminatory. They have a strong socio-economic class and sex bias, and they favor middle-class society and norms at the expense of poor children and children from cultures differing from

the majority, middle-class, Anglo-Saxon culture of the United States. A blatant example of the sex discrimination in standardized achievement tests is illustrated by a question in a primary level MAT which shows an outline drawing of a man in a long coat,with a small mirror attached to his forehead, examining a child. The correct answer for the work that describes the person pictured must be chosen from four choices that include both "doctor" and "nurse." This not only encourages the stereotype that doctors are usually men, it penalizes a child who knows that male nurses exist.

The discriminatory nature of the tests towards other cultures is evident on inspection. On the whole, the tests show white, middle-class children performing stereotyped activities which can be recognized by conventional symbols and the language used to describe them. Strong evidence for the confusing and limited nature of the tests is found in a pamphlet by Deborah Meier (1973), a teacher in New York City who discussed the tests with children in school. She found that the children were confused by the questions, by the unfamiliar language used, and by the situations depicted which were not appropriate to their experiences in life. For example, a question on a primary MAT shows a smiling girl carrying some books in the rain. The correct answer, to be chosen from one of the three sentences that describes the picture, is "Mary's books will get wet in the rain." But, the New York City children argued, this could not be the right answer. She would not be smiling if her books were going to get wet; so they chose things like "the rain will not hurt the books" or "Mary is taking good care of her books," the two other possibilities. Many examples can be chosen, the point being that a significant number of items on nationally used standardized tests are confusing, and choosing the correct answer depends not on reading ability alone (which the tests are supposed to measure), but on knowledge and acceptance of cultural norms.

2. Standardized tests simply are inappropriate for whole categories of educational settings. The very nature of the tests, the way they are given, the way they are graded, and the way their results are used is antithetical to more cooperative open styles of education. This point has been carefully made by Margaret deRivera (1973). She lists a series of ways in which the test situation itself is incompatible with open education practice:

1. *Open classroom*: Children are encouraged or at least allowed to share, to converse, to help one another.
Testing situation: no talking, no sharing, no helping one another.
2. *Open classroom*: Children exercise and demonstrate their knowledge and skills in many different modes: verbally, by action, dramatics, writing, etc.
Testing situation: the children's response mode

is limited to reading, listening, and marking. Knowledge and skills which they are used to exercising in one mode have to be translated to the mode of response that fits the test.

3. *Open classroom*: generally flexibility is such that children can finish most tasks they begin and can go on to something else when finished. Children can move around the room.
Testing situation: no moving on to the next task when finished, often not enough time to finish a task. Children must remain seated at a desk.

4. *Open classroom*: children generally work at many different tasks, so that comparisons are not easy and competition is not encouraged.
Testing situation: children work on the same task at the same time so that comparisons are facilitated.

5. *Open classroom*: each child is viewed as a complex, unique individual, having strengths and weaknesses but essentially qualitatively different from others.
Testing situation: quantitative differences between children are important, qualitative differences are lost. Success is defined by others' failures. (The 60th percentile means that 60 percent of the children in that grade score below.)

6. *Open classroom*: the child is given learning experiences designed to develop a self-image of a competent, effective, successful person. This is considered an important attitude for effective learning.
Testing situation: the very children (those who are weakest in skills) who need the support of a positive self-image in order to continue learning, are discouraged and frustrated by failure.

7. *Open classroom*: thoughtful, critical thinking is encouraged.
Testing situation: often random guessing is a more successful strategy than thoughtfulness since the tests are limited in time. Thoughtfulness is not rewarded.

8. *Open classroom*: intrinsic motivation (i.e. learning for learning's sake) is considered the most effective motivation for long-term learning.
Testing situation: extrinsic motivation (i.e. learning for some outside reward) is encouraged; learning in order to pass the test.

3. Some serious questions are inherent in the methodology used to prepare standardized tests. I have already discussed some of the general implications of the experimental methods on which standardized tests are based, but there are even more detailed problems associated with them. The standardized tests in use in the United States today are prepared in such a way that they are 'valid',

that they will give a 'normal' distribution of results, and that they represent the most common curricula in use. Each of these concepts has serious problems. By 'validity' the test makers mean that the results on the standardized test have been correlated with results from some other measurement. But, in fact, reading tests are not correlated to some independent measure of the ability to read: the correlation that is generally used is only to other grades or tests in school. They are correlated to other paper and pencil tests, usually of the intelligence or achievement kind.

The tests are also constructed to show a 'normal' distribution of children, one smooth curve with not too many spread out at the bottom and not too many spread out at the top, and most of the population distributed around some average value. Two main arguments are used to justify this procedure. First, it is argued that this is generally the way attributes distribute themselves in any large experimental population: if you measure the height of many children of one age, you will find a 'normal' distribution, with a large number of children near one particular measurement (on both sides of it), and the rest of the population trailing off to much greater or lesser heights. Whether this holds for the entire population in such developmental activities as reading is not known and there is really no way to find out.

There is something quite arbitrary in the notion that at every age and every developmental level, no matter what property is tested, the results will distribute evenly along a normal distribution curve; that is, some people cannot do it, some can do it quite well, and the majority does it adequately. Certainly, if a number of 18-month to two-year-old children were tested to see how many steps they could walk in a fairly straight line, the population would distribute itself more or less normally, with some children not being able to walk at all, and most of them only able to manage a small number of steps. (Of course, even here the distribution would not be normal, because a few children might walk so well that the measurement of individual steps would be almost silly.) But a test of ability to walk at age six should yield something quite different from a normal distribution. First of all, we would expect all children, except a small fraction of handicapped children, to be able to do the activity. Then, to set up a walking test for six-year-old children that would result in a normal distribution would mean, first, a strange definition of "walking," and secondly, deliberately devising test items (such as walking on your hands, or running fast or doing complicated dance steps) so that the *nature of the test* would force a normal distribution of the results. This is precisely the situation with reading tests. They are constructed at every level from pre-kindergarten to high school so that the population that is tested will distribute around some norm.

The second justification that test makers give for using normal distributions is that the statistics and the methodology for such distributions are well known, and easy to work out. But even normal distributions, if that is what the population shows, can have variations. The horizontal shape of the curve is important: do most people cluster around a mean, with only a small fraction of the population trailing off at the extremes, or is there a very wide spread of results with only a slight cluster around the mean? Tests are constructed with some spread determined that will make the grades and scores easy to handle: not too much spread and not too little. This characteristic is particularly important when the test is given to a population of students who generally either don't do very well on the test or do extremely well: the standardized tests tell you mainly that you can't say very much about these children from that particular test. But, of course, in education it is precisely the children who are far from the average about whom we need the information.

There are also some questions about the standardization methods of the tests. The problem of finding a test population of children to standardize test items is really quite serious. The 1958 version of the MAT was reported to be standardized against a sample that greatly over-represented southern and rural school districts at the expense of northern and urban districts (Hunter and Rogers, 1967). In order to evaluate a test item, someone or some group of people must go into schools, find thousands of children, give them the sample test, and see what fraction of the children get the correct answer. Now anyone who has worked in schools knows that gaining entry to classrooms to do any sort of research or study is not a random process. It involves a certain amount of political work, getting to know school system people, and choosing school systems and individuals who are cooperative. School officials, quite reasonably, want to know where strangers go and what they do. So the work that must be carried out to standardize a test already raises questions about the nature of the sample.

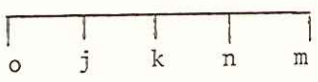
Further, to obtain an appropriate body of questions, the test makers not only average and manipulate the difficulty of the questions, they also design the content so that it will reflect the most widely used curricula. And, since shrewd publishers develop curricula with an eye to matching the tests, that is, contain the most-used words, etc., a vicious cycle ensues in which the tests and curricula (developed by the same groups) justify each other, while having little relation to the lives and achievements of children. Any examination of tests will reveal that the vocabulary, style, and material content are very much school-oriented, and not life-oriented. They certainly do not contain any vocabulary or structure corresponding to black English, as described by Labov and others (Labov, 1972). But neither do they really contain the language of any children. The test words and stories are

a bland melange of the dull fare found in school readers. One even looks in vain for evidence of the newer curricula that have been introduced into the schools. It is widely assumed, for example, that the 'new math' has taken over the schools, that set theory, other-than-base ten system, and various mathematical definitions have become important. This certainly doesn't show up in the tests. As members of Educational Developmental Center's (EDC) Project One have shown in a recent analysis of the math tests, 50 to 70 percent of the questions deal with simple computation in the base ten system and the rest of the material is heavily directed towards simple definitions. The few questions that deal with modern mathematical concepts are often ambiguous or misleading, and sometimes just wrong.

4. Hierarchy of Knowledge. The last three points all deal with consequences of assumptions inherent in the development of the tests, rather than with their general characteristics. The process of constructing test items --definitions, problems, words--proceeds under the assumption that there is a clear hierarchy of knowledge: that some things are harder than others, that some activities are, and should be, learned later than others, that the kind of problems that children can solve or the kinds of material they can read can be strictly graded and categorized from simple to complex. This assumption runs counter to several important principles of learning theory supported by open education practitioners. I have already discussed these individual differences: learning styles, horizontal growth, and individual rates of development.

5. Standardized tests used in the United States today are exclusively paper and pencil tests which measure nothing but simple reading skills, the naming of concepts or objects, and computation skills. Despite the titles to the sections of the tests, very little else is measured. Most reading tests have a section entitled "Comprehension." But one way to answer the questions is not to read a paragraph and comprehend it, but simply to skim the paragraph, look at the questions, and then find the salient information. The tests certainly do not measure the comprehension of ideas; at most, they may determine whether the person taking the test knows the meaning of a word. The math sections have such titles as "Concepts" or "Problem Solving," but the concepts usually are definitions or names and the problem solving is more often a reading problem than anything else.

6. The fact is that the standardized tests that are given are just plain bad. They are not even good tests by their own standards. For example, the Primary Form F of the MAT shows the children a math problem with the following figure:



A child is asked to "Look at the line segment at the top of the box. Fill in the space next to the statement which is true."

The statements are: *k is greater than n*
m is less than j
j is equal to k
j is less than k
dk

It doesn't take too much mathematical knowledge to know that you cannot define a line segment by one point. The statements are meaningless.

I have picked this one example because it is not just a case of the question being vague, ambiguous, or misleading. The question is simply impossible to answer at all. It may seem like a small matter that one out of a set of 40 questions in a test which has a total of 114 items is incorrect, but in fact the consequences of an impossible question are quite significant; one question can make a surprisingly large difference in a grade equivalent score. But a more important point is that these incorrect questions, as well as many more that are ambiguous and strange, appear on the tests at all.

7. Probably the most specious argument made in support of standardized tests is that evaluation is too important an activity to be left to individual teachers and schools, and to the dangers of a great variety of standards and a good deal of sloppy measurement. Everyone knows how hard it is to make up good exam questions, the argument goes, so better leave the process to the 'experts' who test out the questions on large sample populations and ponder them carefully.

But the experts seem to come up with grossly inadequate measures. My first contact with the world of standardized testing was as a chemistry teacher in a private high school. I had a very bright, small class and we worked hard. Many of the students were the children of Caltech faculty, they were interested in science and had good training. At the end of the year, I gave them a standardized examination prepared by the ubiquitous ETS, organized especially for independent schools. (As far as I know this test is still being given.) But the test, I found, contained some questions that were simply incorrect: a drawing of a laboratory experiment showed a totally unsafe situation which might blow up at any moment, and some that were simply irrelevant: What is the Solvay Process? The latter was, in fact, an industrial process already becoming obsolete at that time. In my youthful enthusiasm and anger, I showed the test to a number of faculty members--prestigious chemists, members of the National Academy of Sciences, and leaders in their field. They all agreed that the test was stupid, wrong, ambiguous, and inappropriate for a reasonable chemical education. Yet when I wrote to ETS about it, I got the same answers that the supporters of tests still give: They also had consulted experts who saw nothing wrong with the test, they had gone to a good deal of trouble to standar-

dize the test questions, and they simply couldn't go about changing them.

WHY ARE THE TESTS SO BAD?

But the major concern about the tests and their influence does not depend on the particular criticisms that can be levelled against them. The tests fail by the very standards of the experimental paradigm within which they are made; that is, they are poor tests with ambiguous and incorrect questions. What is of greater concern is the way the tests succeed within the wider framework in which they are used: namely, they are one more component in the sorting system of American schools. They contribute one element (although not the only one), one necessary condition (although not a sufficient one) to see to it that the schools continue the society as it is. Society uses schools to sort out and classify, to reward those who come from the middle-class and keep down those who are already poor; and the tests help in this major social effort. They couldn't do it alone, they simply contribute. And as long as they do that job, which happens to be independent of the specific test items, unrelated to whether or not there are ambiguous questions, they can continue to be used and used effectively (Karier, 1972).

The research paradigm within which the tests are constructed is actually very good for determining major trends, making gross distinctions: distinguishing between those who can read in general and those who cannot, between those who can compute reasonably and those who really struggle with numbers. This sort of distinction is easy enough to make, and since the test design is good enough to determine these gross differences it doesn't really matter too much if a few questions are ambiguous. Actually the ambiguous questions also serve an important function: they make the tests better at the kind of classifying for which they are used. The tests don't do very well at describing individual styles, levels of achievement, or usable knowledge, but they do test the ability to follow instructions, to not think too deeply (that's one way to avoid the ambiguities in many questions), and to do reasonably neat clerical work at a steady pace without thinking about it too much.

One measure of the extent to which the tests don't accurately reflect the abilities and knowledge of individuals is the number of exceptions to expected results. Every person active in education has her own store of anecdotes about Jane who did poorly on an MAT, but could do the work; of Frankie who could read only on the second grade level, but after two months of help could read on the sixth grade level; of Janice whose IQ rose 25 points in a year. In some cases, where people have looked carefully at children and worked sensitively with them, whole classes and groups have made phenomenal increases in their IQ scores or their grade level achievement over relatively

short periods of time. In *Reading, How To* (1973), Kohl reports the case of Lillian, a child whose performance improved so much that it required the threat of a law suit to force the school to accept the results of three reading tests. This phenomenon is further documented by a report from the Far West Laboratory for Educational Research and Development (Rayder and Nimnicht, 1973) concerning some of the results in their Follow Through program classes. The authors demonstrated that the children in classes in 14 school systems across the country increased their *average* IQ scores on the Wechsler test of intelligence by significant amounts over a three-year period of the program. They went from scores that were much below the average for the country to scores that were above that norm. The authors concluded:

> First, intelligence tests are not reliable measures of the abilities of these children....second, the problem of cumulative deficits is with the school not the child.

In other words, standardized tests are one link in a long process that tells poor children, and especially poor black children, that they are on the bottom of the heap and should stay there. That is why American schools continue to use tests which are inadequate even by their own stated goals, and which have become one of the principle instruments through which schools serve to maintain social and economic inequality.

8

Evaluation Alternatives

REFORM OF STANDARDIZED TESTS

One approach to "*the* major disaster area in education," as evaluation was recently called by James B. Macdonald (1974), would be to improve the standardized tests. From the foregoing criticism, it is obvious there is room for a great deal of improvement. The questions could be better, the standardization could be more representative, and the validation against criteria more appropriate than the ones that are used. More imaginative use of the available technology could vastly improve even paper and pencil, machine-graded examinations. If it is accepted that there is more than one way of doing a problem, why not present the alternative ways on the test and grade anyone 'correct' who, simply, solves the problem, whichever way he or she does it? The whole notion that the scoring and administration of the MAT is done on a basis of total correct answers in each area without any further modification is really quite absurd. Why not a choice of questions, or questions which relate to a wider range of skill, or the possibility of more than one correct answer in some cases? Moreover, is there any reason at all to limit the concept of standardized achievement to paper and pencil tests? Why not standardize a much broader range of activities if this were desired?

Unfortunately, any effort to reform the tests has two major drawbacks. First, it ignores the analysis of why the tests are so bad now. To assume that achieving better standardized tests is simply a matter of making changes in the tests themselves is, I believe, to be naive about the education world and about American society. It is highly unlikely that all the people who put the tests together, suggest the questions, write the language, try them out on children, standardize them, and finally publish and sell them are all totally unperceptive and uneducated. The tests and their use are deeply embedded in the fabric of American society and must be rejected on political grounds, not modified at the technical level.

Secondly, any proposal for a major effort to produce new testing mechanisms is reminiscent of the program that was launched almost 20 years ago to produce new science and math curricula. Scientists and mathematicians who turned their attention to schools were horrified at the state of the situation: the curriculum was simply bad,

they said, full of error, wrong concepts, incorrect statements, too much stress on rote learning, simple drill, etc. They set out to reform education by updating and correcting the curriculum, to make it 'better'. One of the major learning experiences for those involved in that curriculum reform was that new curricula, although a necessary condition for better school experiences for children, was hardly a sufficient change. In fact, much of the new curriculum was neatly fitted into existing school structures (indeed it was designed for this) and instead of the curricula changing the schools, the schools absorbed the new curricula without much modification in the essence of the schooling provided for most children. In many ways the new curricula was simply ignored. While the rhetoric of the New Math has had wide acceptance in the schools it would be hard to know it from many of the day-to-day activities in the classrooms, and difficult to discern it on the items which appear on the standardized tests (Sarason, 1971).

To try to 'correct' or save education by simply outfitting the schools with better testing procedures is inadequate as a strategy. As in so much else, parts cannot easily be separated from the whole. To bring about fundamental change in the schools, the entire program must be reexamined: curriculum, evaluation, teaching style, views of learning and knowledge, etc.

There is obviously some merit in developing a more reasonable and wider-ranging approach to standardized testing, as long as one neither expects the task to be simple, nor hopes to change education by this alone. The area of developing alternative tests is a wide-open field; remarkably little work has been done in it because the standardized achievement tests and their companions, the widely used intelligence tests, so dominate the field that little else has been tried and certainly little else has been carried vary far. An appropriate analogy can be made with the automobile industry. At one point, in the early development of automobiles in the United States, a wide range of design and approaches to the problem of mechanical energy-driven vehicles were explored: different engines (electric and steam) as well as other fossil fuel (such as diesel fuel) competed with the high-octane gasoline model. But the gasoline-powered internal combustion engine was so successful, it spread so widely over the market, that many other technologies were simply not followed up very much. Today we know a great deal about the gasoline engine that uses rather a lot of gasoline, and very little about the alternatives. Its commercial success and relatively low cost (which was related to that success), along with the low value placed on the various problems it represented (that is, as long as there was no gas shortage), simply made it unnecessary to do other work.

There is, however, another component of this analogy which is not quite so innocent. Along with developing its technology, the automobile industry evolved policies that channeled and directed research, labor, and expendi-

tures in the direction of private automobile travel and away from mass transit. Decisions that had profound effects on our society served to benefit a particular sector of private industry, namely the sponor of those decisions. As the *Boston Globe* observed in commenting about a recent Senate subcommittee report:

> GM, Ford, and Chrysler reshaped American ground transportation to serve corporate wants instead of social needs. This study suggests that a monopoly in ground vehicle production has led inevitably to a breakdown on the nation's ground transportation.
>
> The report further documents how, beginning in the 1920s, General Motors began to buy up rail and electric urban transportation systems and then replaced them with buses or diesel locomotives, which it manufactured (March 10, 1974).

The same report, the *Globe* reported on March 3, 1974, also documents that changes in styling in the automobile industry through the years were not necessarily related to improvements in technology (Rothschild, 1973).

It may well be questioned whether there are similar interests involved in the continuing use of large-scale standardized testing programs in our urban centers. The companies that produce standardized tests are analogous to the big three automobile manufacturers: they dominate their market and dictate what is and isn't profitable, but their outlook is limited by what they have found successful. Commercial self-interest makes them unwilling and unlikely to speculate on different projects that would undercut their own positions. And, like the big three automobile manufacturers, the publishers who produce testing programs are not isolated from the rest of society. They have connections in schools of education, foundations, and government that work together to maintain the *status quo*, just as the automobile industry has connections in research institutes, regulatory agencies, and government.

One strong argument continually made for maintaining the present evaluation system is the cost factors involved. It is simply a great deal cheaper to give the MAT to every child in the school system than it would be to introduce any of the alternatives that have been suggested. It is undeniably correct that it is much cheaper in dollars and cents for any particular school system in 1974 to buy MAT booklets for every child and give these tests than to establish some sort of individual observation system to determine the status of each child. But the total expenses are so different that they cannot be compared because it is a little like comparing the cost of gas for your kitchen stove and the cost of installing a nuclear-powered technique for preparing food. A kitchen that already has a gas stove will also have appropriate cooking utensils, a line leading in for the gas, and stores nearby which sell food that can be easily prepared by gas stoves in a short time. To compare the real costs of two totally dif-

ferent approaches to food preparation, one would have to take into account the investment that has been made in all these things and the development costs that went into setting up a food distribution network to cater to that style of cooking.

The cost of feeding the present testing machine is quite small in comparison to setting up another one, but that does not mean the total investment in it is small. In fact, school systems spend a great deal of money on testing and evaluating children. Besides the cost of the millions of test booklets, which are not reusable, there are a number of personnel in the school system, especially the city systems, but smaller ones as well, whose job is to give the tests, organizing the test-taking, etc. Teachers and children spend a good deal of time giving and taking tests. In some Follow Through sites, as many as six weeks of the spring term were totally lost while the classes went through the agony of taking the various required tests dictated by the city, the program, etc. The whole experience simply disrupted all instructional activities for a month and a half (that is about 18 percent of the *total* school year). Nor do the above costs include the human and social factors: how the tests affect programs, how they tyrannize teachers and demoralize students. Also not included is the incredible inefficiency of testing. Typically, children are tested sometime in the fall and spring and the comparative results are released very late in that year or, often, in the next year. Teachers cannot even use the tests for their own teaching purposes; they can only be used as a weapon by outsiders, after the children have moved on to the next grade.

ALTERNATIVE STRATEGIES FOR MEASURING CHILDREN'S LEARNING

In terms that have been made familiar by Thomas Kuhn (1970), there is always a prevalent *paradigm* in any scientific activity (perhaps in any human activity) within which a majority of the work is carried out. But there is usually a small minority of work going on outside it, and the major breakthroughs in science occur when a new paradigm replaces an old one. Likewise, in evaluation work, the vast majority of activity falls within the accepted experimental-psychology-research paradigm, but there has been a small ongoing tradition of work outside that paradigm, and open educators are waiting hopefully for the over-throw which will allow a breakthrough in our views on evaluation. There are indications that evaluation alternatives are becoming more popular (Eisner, 1972; Parlett and Hamilton, 1972, etc.).

An older American evaluation effort (Aikin, 1942) is worth discussing briefly because it transcends the paradigm. In 1932, the Progressive Education Association launched a major effort to determine what, if any, influence progressive education practice had on students. A large group of students from 30 high schools scattered

around the country were followed throughout high school and college for a total of eight years. The evaluation activity involved a number of factors besides standardized measures on students. The staffs of participating high schools were particularly concerned about their programs during this time and used the fact that they were part of the study to examine and modify their activities, and the colleges involved agreed to waive admission standards for the students involved. The study led to meetings between the cooperating schools and colleges, and it stimulated curriculum changes in both.

The actual evaluation work included questionnaires, records, unobtrusive measures, interviews, etc. The best description of the evaluation/education activity can be obtained from quoting the summary of their neglected five-volume work:

> In the comparison of the 1,475 matched pairs, the college Follow-up staff found that the graduates of the Thirty Schools
>
> 1. earned a slightly higher total grade average;
> 2. earned higher grade averages in all subject fields except foreign languages;
> 3. specialized in the same academic fields as did the comparison students;
> 4. did not differ from the comparison group in the number of times they were placed on probation;
> 5. received slightly more academic honors in each year;
> 6. were more often judged to possess a high degree of intellectual curiosity and drive;
> 7. were more often judged to be precise, systematic, and objective in their thinking;
> 8. were more often judged to have developed clear or well-formulated ideas concerning the meaning of education--especially in the first two years in college;
> 9. more often demonstrated a high degree of resourcefulness in meeting new situations;
> 10. did not differ from the comparison group in ability to plan their time effectively;
> 11. had about the same problems of adjustment as the comparison group, but approached their solution with greater effectiveness;
> 12. participated somewhat more frequently, and more often enjoyed appreciative experiences, in the arts;
> 13. participated more in all organized student groups except religious and "service" activities;
> 14. earned in each college year a higher percentage of non-academic honors (officership in organizations, election to managerial societies, athletic insignia, leading roles in dramatic and musical presentations);

> The graduates of the most experimental schools were strikingly more successful than their matches. Differences in their favor were much greater than the differences between the total Thirty Schools and their comparison group. For these students the differences were smaller and less consistent than the total Thirty Schools and their comparison group. (p. 148)

Other work has been carried on in an effort to develop evaluation alternatives. These efforts can be classified as follows:

1. Different 'Standardized' Tests. One reform which has been proposed is to move from 'norm'-referenced tests to 'criterion'-referenced tests. In criterion-referenced tests, items are not correlated with some other scale of what children do on these tests or with some standardization which simply compares children with each other. Instead items are correlated with actual ability to carry out some task. A norm-standardized test can tell you where a child stands relative to the rest of the population that was used for the normalizing procedure on that test item; a criterion-referenced test can tell whether a child can do something that has been correlated with that item. In principle this sounds quite good, and, in fact, if carefully done, it can lead to a much more satisfactory approach to testing strategies. But there are some serious difficulties. The ultimate in criterion-referenced tests is doing the task itself. If you want to know whether a student can repair a car, you have her repair the car. But, of course, the whole idea of standardized tests is to substitute some simple easily reproducable and generalizable activity for the things you really want to test for. The more complex the activity that you want to evaluate,the harder it is to make a reliable criterion-referenced test. This is reflected in the fact that many tests that are reported to be criterion-referenced leave some question about the relation between what is tested and the activity, or, more commonly, have defined a trivial activity,or one that only has reality in the world of tests,as the criterion that has been used as a reference.

It has long been a standard procedure to have 'lab' exams in experimental science subjects. Many biology students remember vividly the difference between recognizing a drawing of a microscopic object on a paper and pencil test and identifying it under the microscope in a practical exam. Much of the knowledge that children gain in school is of the practical, hands-on type, and could be tested accordingly. It is particularly inexcusable that science learning is evaluated almost exclusively by paper and pencil tests which essentially measure reading ability and little more. Even the definitions that are so prevalent on the science portions of standardized tests usually measure only two things--whether the student can read the name of some scientific object or principle, and whether the student can associate that with a related term. Neither of these skills

covers a significant fraction of what could be considered scientific literacy. Also, the line drawings which accompany many test items for younger children are only a partial substitute for naming; they are highly stylized and symbolic representations, not even photographs.

Some research groups have substituted objects, photographs, diagrams, and manipulative materials for paper and pencil test problems. This makes it possible to discover a number of things about children's abilities independent of their reading skills. First, a child who understands the principle of an electric circuit can light a bulb if given the proper materials even if she could not answer a written question about the subject. Secondly, using materials tells you something about the way a child goes about a problem. Are groups of objects simply enumerated or are sub groups added or multiplied? The actual way a child manipulates material informs the observer about the approach used much more than any particular answer on a scored sheet. Most people who bother to do this kind of work with children usually come away profoundly impressed with the limited notions they have of how children think and learn. This type of problem is just as objective as any paper and pencil test, or at least it can be made just as objective.

In a recent ambitious evaluation effort (Comber and Keeves, 1973), hundreds of thousands of students in 19 countries were given extensive standardized science tests in order to assess science education on a global scale. The extensive technical document which reports the results, all based on paper and pencil tests that required considerable reading ability, contains the following tantalizing comment:

> Perhaps of special interest, in view of the current debate on the place to be accorded practical work in various kinds of school science, was the attempt to produce optional tests of practical abilities requiring only very simple and easily obtainable materials. Unfortunately, only two countries elected to take these 'practical' tests, but the evidence from these suggests that such practical tests measure quite different abilities from those assessed by the more traditional tests, even those designed to assess practical skills as far as possible without resort to actual apparatus. It follows that if students' firsthand experience is to become an essential feature of school science, as many Science teachers believe it should, then the further development of such tests will be highly desirable, if not imperative. (p. 288)

2. Materials can be used to make possible open-ended forms of evaluation by not determining beforehand what question is to be asked of them. Such an evaluation was performed by Eleanor Duckworth (1970) for the African Primary Science Program. In trying to find out whether exploration,

discovery, and work with a wide range of materials had any appreciable effect on the children, she carried out a study in which she took two groups of children: those who had had exposure to the APSP course (a materials rich, manipulative science program) and those who had had only traditional education in school. She simply placed them in a room with lots of material (not the same materials used in the APSP courses), and watched what happened. She noted that the test group--those who had been exposed to APSP--were more inquisitive, did more things, more connected and sequential things, asked more questions of their environment and used it more adeptly than a group of children who had not been so exposed. It should be relatively easy to extend this approach to evaluation to the day-to-day life of American schools.

In this approach, the observer isn't certain before doing the work just what behavior will occur in the experimental children. It is an open-ended evaluation: an effort to say, "let's see what these children do." In this sense, it is an application of the most sensitive and sensible evaluation strategy of any one of a number of activities based on the approach of the 'clinical interview'. Obviously, the only way that we can ever measure the new or novel things that children do is to have an assessment instrument that leaves room for observing new and unexpected behavior. This requires both the input of enough material from the observer to give the child something to work on, and enough freedom on the part of the respondent to take advantage of it. The style represented by the Piagetian interview of finding out 'where children are at' is perfect for this approach. Using this same approach it is also possible to find out where groups of children are with respect to certain concepts, or types of problems, or styles of knowledge.

Deborah Meier's revealing study about children's responses to the MAT is an example of the use of a clinical interview to find out what children know. In this case, the material of the evaluation was the standardized tests which the children worked on. By talking with them, it was possible to find out a great deal about their knowledge, assumptions, frames of reference, etc.

3. Check lists for teachers to guide them in evaluating children's learning are powerful evaluative tools. Some lists are available to cover reading achievement, math skills, and science knowledge. Lists of this sort have a tremendous flexibility of use (although they are also subject to the danger of overly rigid application), they do not require elaborate test administration procedures, they can be individually applied and they provide information directly to the teacher. One big difference between check lists and more formal tests is that they usually are not considered total descriptions, but guides. In fact, if they get too detailed, they become less useful. A list of reading accomplishments need not cover every technical detail of a child's reading mastery, but it will give a teacher a sense of where that child has

arrived at and what the child needs help on. At the same time, it serves to remind a teacher of skills or parts of a process that may be missing from a child's repetoire. An example of such a diagnostic, open-ended reading guide is presented in *Evaluation Reconsidered* (Norris, 1973).

4. Record Keeping. A classic means for evaluating children's growth and development is some systematic recording of facts or events that involve them. This is the basis of any sensible evaluation of children. It is a method that all parents use informally. We observe our children, note the changes they undergo, and judge their development on the basis of these changes. It is fairly easy to note major differences with a small number of children, so most parents don't keep records of when their children first walk or talk or perform certain intellectual feats. In a school, where there are more children per adult and the adults concerned with the children change from year to year, more formal records are necessary. The problem is that most schools keep rather dull, and not very useful records: most often some adult assessment of the general level of the child and a compilation of standardized test scores. The anecdotal records are usually spotty and incomplete, while the standardized reading scores are simply not helpful, even on a cumulative basis.

Used more imaginatively, record keeping has vast possibilities for assessing the growth of children.* The work of the Bureau of Educational Experiments, founded in 1911 (recently reprinted), contains explicit discussion of efforts to assess children's growth through record keeping before World War I (Winsor, 1973). In the pioneering reform movement in the Vienna school system between the World Wars, report cards were abandoned and, instead, each child was given an elaborate form which recorded aspects of her social, intellectual, and emotional development (Papaneck, 1962).

*See also Engel, B., *A Handbook on Documentation*, in this series.

A more contemporary extensive and thoughtful effort of documenting children's growth and development has been carried out for nearly a decade by Pat Carini (1973) of the Prospect School, North Bennington, Vermont.* By keeping a variety of records, she and her colleagues have amassed an impressive amount of revealing information both about general aspects of children's growth and specific information which is helpful about particular children. Included among these are:

*See also Carini, P.F., *Observation and Description: An Alternative Methodology for the Investigation of Human Phenomena*, in this series.

Children's work: e.g., drawings, photos, etc.
Children's journals (generally only for children aged 11 and older)
Children's notebooks and written work
Teacher's weekly records
Teacher's reports to parents
Teacher's assessment of children's work in math, reading, activities
Curriculum trees
Sociograms

Records are another 'objective' form of evaluation, and the longer they are kept, the more objective they become. A single estimate of how much time a child spends in math activities may be way off, but 10 such estimates in a month probably average out fairly close to a correct figure. One component of any successful record-keeping activity is longevity. Almost any measure or record becomes interesting and able to tell you something if you keep it long enough. Historians have long ago learned the power of such apparently 'trivial' data as vital statistics when available over long periods of time.

Of course, the establishment of a record-keeping system is not an easy task. Who does the work, who stores them, who looks at them, what do you record, when, how, etc.? All these are questions that have to be addressed; then someone has to see to it that whatever procedure is adopted is maintained consistently for long enough so that information can be drawn from it. But this sort of evaluation has proven to be an extremely useful way to know what children are doing, what they are capable of, and the areas in which they need help. Records also provide invaluable information for program evaluations.

EVALUATION AT THE PROGRAM LEVEL

The whole field of evaluation is much larger than the concern for the evaluation of individual children's growth and development. To the extent that the present methods used in the schools to measure children's achievement are inadequate, this inadequacy is magnified at all other levels of evaluation. The public schools simply do not have thorough unbiased methods developed within their setting for systematically knowing and recording children's development and progress, and what the next best steps for them might be. Also, the public schools have not developed adquate systems to support teachers making day-to-day decisions about the best opportunities to provide for children. The present system, with its tabulations and aura of objectivity, simply permits administrators to feel they know what is happening and can make rational decisions. A number of schools follow the barbarous custom of posting the standardized achievement test scores in the principal's office by grade and teacher, so that the teachers can all be compared in terms of the results and so that, presumably, they will have an incentive to 'raise' the standing of their class. It is certainly the case in many schools that teachers believe, with good reason, that their future salary increments and promotions depend on these results. The test system therefore becomes yet another competitive situation in the schools, with higher scores becoming the production goal, like Stakanovite practices in Russian factories under Stalin.

Yet is it not usual for descriptions of programs, statements of educational aims, and official instructions to personnel to include broader goals than simply the

attainment of certain scores on children's achievement tests? These goals may cover affective growth, social situations, interest, personal growth, and a wide range of other issues. If these larger issues are taken seriously, then a wider range of evaluation strategies must be employed. Where this imperative has been recognized,every conceivable activity has been used at one time or another to assist in making judgments, including interviews, questionaires, other psychometric tests, cost analyses, communities' reactions, hunches and political considerations. Because the range of activities that may be involved in the wider range of evaluation situations is so broad, no specific critique is possible.

Often the political situation is such that even though the funds available are not sufficient for a thorough analysis,a 'formal' evaluation must be carried out. As there is no easy approach, some hodgepodge of activity is thrown together and called evaluation. It is in these instances that it becomes transparently clear that the so-called objective evaluation is preoccupied more with political and social issues than methodological ones. Of primary concern are questions about who wants particular programs, about what their benefits are on the basis of broad social terms, about what people have to gain or lose by the implementation of a program or by the hiring of a teacher or of a superintendent, etc. This is not to deny that a considerable body of data, measurement, and material can be relevant to decision making and should be gathered and used as much as possible. Rather, it is to say that there are no totally objective approaches to decision making, as it involves people's most basic beliefs, prejudices, and feelings.

In summary, to improve the situation of evaluation in American schools, two things need to be accomplished. First, the scope of what is considered evaluation has to be vastly broadened, and this work has to become an integral part of the educational experience. Evaluation is judgment, and to make judgments the relevant information must be assembled. It is foolish to limit what is measured and recorded about children or programs to those few bits of data that happen to be available from present standardized achievement tests. If evaluation is looked at from the point of its relation to the rest of the educational program, one can recognize how separate the two are at present. It becomes especially clear that children are hurt and discouraged by the present system, while teachers are simply not assisted in their difficult tasks.

Secondly, judgments of evaluation are part of an all-encompassing political-social atmosphere. One cannot expect that the formal part of evaluation will deviate very far from the more general, informal judgments that are meted out by the overall society. If the society decides that black children are not as worthy as white children, or that girls are inferior to boys, then the formal evaluation system will either reflect this judgment or its results will be ignored. We can only hope to bring

about major changes in the ways in which evaluation is carried out at the same time that we bring about major changes in the structure of education and in the society as a whole.

Bibliography

AERA Monograph Series on Curriculum Evaluation, 1. Chicago: Rand McNally, 1967.

Aikin, W.M., *The Story of the Eight Year Study*. New York: McGraw-Hill, 1942, p. 148.

Allen, V.F., *What Does a Reading Test Test?* Philadelphia: Temple University TTT project monograph series, 1974.

Berg, I., *Education and Jobs: The Great Training Robbery*. New York: Praeger, 1970.

Bussis, A. and Chittenden, E., "The Horizontal Dimension of Learning" in A. Tobier, ed. *Evaluation Reconsidered*. New York: Workshop Center for Open Education, 1973, p. 8.

Campbell, D.T. and Stanley, V.C., *Experimental and Quasi-Experimental Design*. Chicago: Rand McNally, reprinted from *Handbook of Research on Teaching*, A.E.R.A., 1963.

Carini, P., "The Prospect School: Taking Account of Progress," *Childhood Education, 49*, 350 (1973).

Children and Their Primary Schools (The Plowden Report). London: H.M.S.O., 1967. Vol. 2, p. 260.

Comber, L.C. and Keeves, J.P., *Science Education in Nineteen Countries*. New York: Halsted Press, 1973, p. 288.

Cremin, L., *The Transformation of the School*. New York: Knopf, 1961.

DeRivera, M., "Academic Achievement Tests and the Survival of Open Education," *E.D.C. News*, No. 2, p. 7, 1973.

Duckworth, E., "Evaluation of the African Primary Science Program," Newton, Mass., Education Development Center, 1970.

Eisner, E.W., "Emerging Models for Educational Evaluation," *School Review, 80,* 573 (1972); Parlett, M. and Hamilton, D., "Evaluation as Illuminator," Occasional Paper No. 9. Edinburgh: Center for Research in the Educational Sciences, 1972; Tobier, A., ed., *Evaluation Reconsidered.* New York: Workshop Center for Open Education, 1973; Sarason, *op. cit.*

Evaluation of the Office of Economic Opportunity's Performance Contracting Experiment, Report to Congress. U.S. Government Printing Office, B-130515, May 8, 1974.

Fisher, R.A., *The Design of Experiments.* London: Oliver and Boyd, 1935.

Goodall, J. van L., *In the Shadow of Man.* London: Collins, 1971.

Halberstam, D., *The Best and the Brightest.* New York: Random House, 1969.

Henry, J., *Culture Against Man.* New York: Random House, 1963.

Herndon, J., *The Way its Spozed to Be.* New York: Simon and Schuster, 1965.

Holt, J., *How Children Fail.* New York: Pittman, 1964.

Hunter, L.B., and Rogers, F.A., "Testing: Politics and Pretense," *Urban Review,* Vol. 2, No. 3, p. 5, December 1967.

Jackson, P.W., *Life in Classrooms.* New York: Holt, Rinehart and Winston, 1968.

Kagan, J., "Cross-cultural Perspectives on Early Development," address at AAAS annual meeting. Washington, D.C., December 26, 1972.

Karier, C.J., "Testing for Order and Control in the Corporate Liberal State," *Educational Theory,* 22, 159-180 (1972).

Kimball, S.T., "An Anthropological View of Social Systems and Learning," in, F.A.S. Ianni and E. Storey, editors, *Cultural Relevance and Educational Issues.* Boston: Little Brown, 1972.

Kohl, H., *36 Children.* New York: New American Library, 1967.

_____. *Reading, How To.* New York: Dutton, 1973, p. 20.

Kozol, J., *Death at an Early Age.* Boston: Houghton Mifflin, 1967.

Kuhn, T., *The Structure of Scientific Revolutions,* 2nd ed. Chicago: U. of Chicago Press, 1970.

Labov, W., *Language in the Inner City.* Philadelphia: Philadelphia U. Press, 1972.

Macdonald, J.B., "An Evaluation of Evaluation," *The Urban Review, 7,* 3 (1974).

Meier, D., *Reading Failure and the Tests.* New York: Workshop Center for Open Education, 1973.

Norris, M., "A Guide for Reading Assessment: Grades 1 and 2," in A. Tobier, editor, *Evaluation Reconsidered.* New York: Workshop Center for Open Education, 1973.

Papeneck, E., *The Austrian School Reform.* New York: Frederick Fell, Inc., 1962.

Rayder, N., Body, B., and Nimnicht, G., "Assessing Follow Through," Far West Laboratory for Educational Research and Development, San Francisco, 1973.

Rist, R.C., *The Urban School.* Cambridge: M.I.T. Press, 1973.

Rothschild,Emma, *Paradise Lost: The Decline of the Auto Industrial Age.* New York: Random House, 1973.

Sarason, S.B., *The Culture of the School and the Problem of Change.* Boston: Allyn and Bacon, 1971, ch.4.

Venezky, R.L., *Testing and Reading.* Urbana, Illinois: National Council of Teachers of English, 1974.

Winsor, C., ed., *Experimental Schools Revisited.* New York: Agathon Press, 1973.

Also available as part of the North Dakota Study Group on Evaluation series:

Observation and Description: An Alternative Methodology for the Investigation of Human Phenomena
Patricia F. Carini

Alternative Evaluation Research Paradigm
Michael Quinn Patton

A Handbook on Documentation
Brenda Engel

Deepening the Questions About Change: Developing the Open Corridor Advisory
Lillian Weber

The Teacher Curriculum Work Center: A Descriptive Study
Sharon Feiman

Single copies $2, from Vito Perrone, CTL
U. of North Dakota, Grand Forks, N.D. 58201