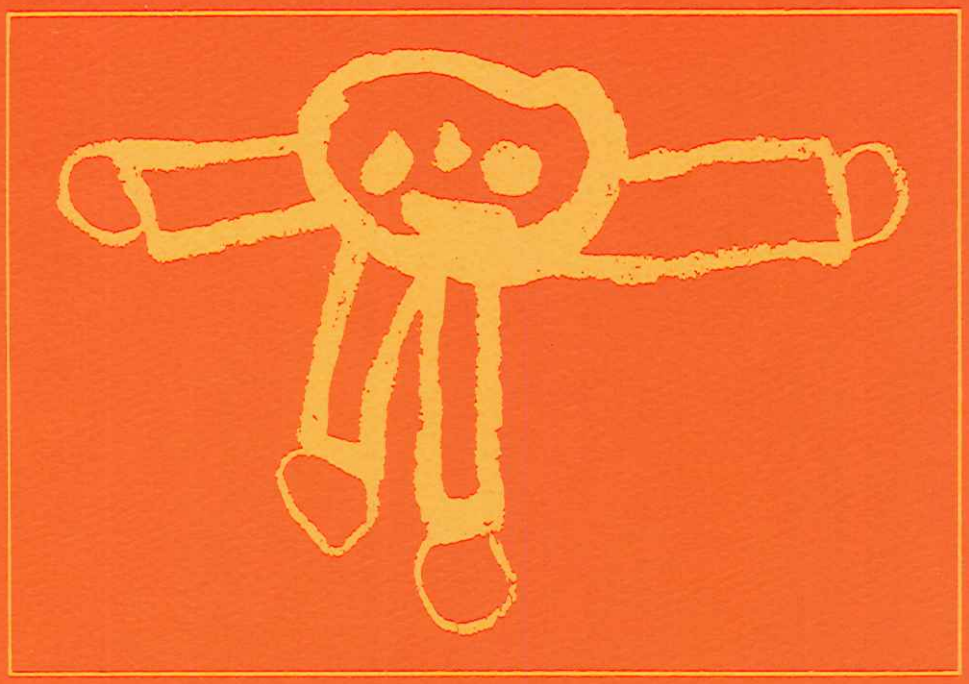
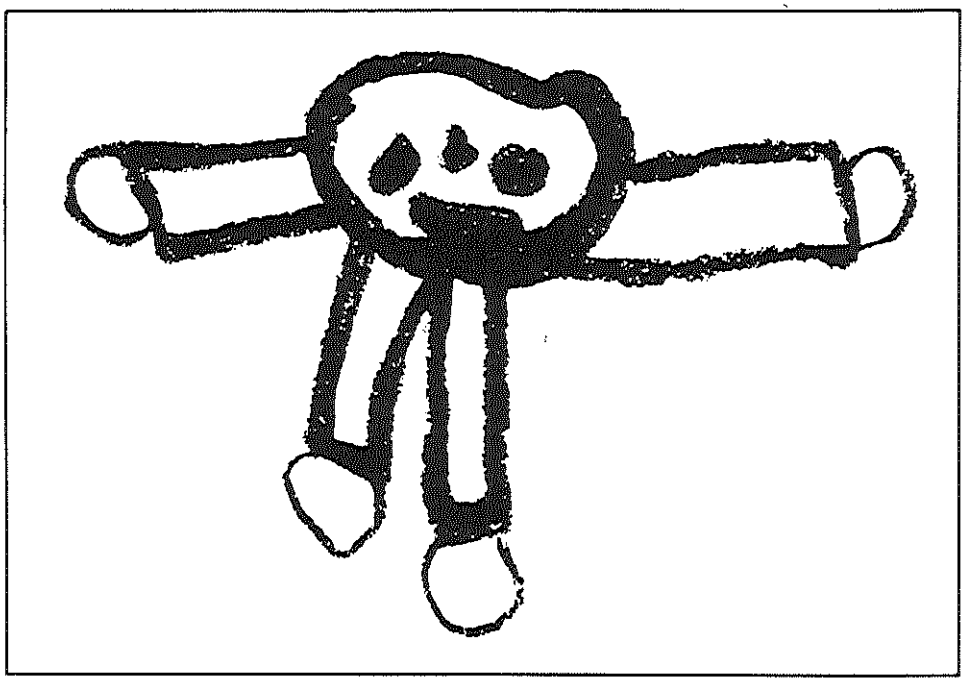
North Dakota Study Group on Evaluation



Michael Quinn Patton

ALTERNATIVE EVALUATION RESEARCH PARADIGM

North Dakota Study Group on Evaluation



Michael Quinn Patton

ALTERNATIVE EVALUATION RESEARCH PARADIGM

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

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: 75-274

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

Introduction

This is a description and analysis of alternative evaluation research paradigms, or more specifically a description and analysis of two contrasting paradigms: one that now dominates the field of evaluation research, practiced by the great majority of academic researchers in education and the social sciences; and another, which in this paper will be referred to as the alternative paradigm, that is rather like an ignored, illegitimate stepchild lurking in the shadow of the dominant paradigm. My purpose in this paper is to examine the alternative assumptions, values, ideology, and perceptions that *inevitably* undergird evaluation research methodology. It is a task whose importance is underscored by the recent explosion of interest in evaluating educational innovations and social action programs.

Part of this interest can be attributed to the budgetary implications of evaluation results, part to the desire of the active public for evaluative information about government programs, part to the needs of program administrators and participants for information about and evaluation of their own programs. As Edward A. Suchman has noted (1967): "The demand that some attempt be made to determine the effectiveness of public service and the social action programs has become increasingly insistent. ...The result has been a sudden awakening of interest in a long-neglected aspect of social research..." Indeed, since 1967, the literature in the field has mushroomed, not only books (e.g., Suchman, 1967; Weiss, 1972a, b; Caro, 1971; Rossi, 1972), but numerous articles in major social science and education research journals. However, what makes consideration of an alternative evaluation research paradigm so pressing is the fact that *these prominent exemplars of evaluation are based on a single, largely unquestioned, scientific paradigm.*

The paradigm, which I refer to in the paper as *The* Scientific Method, derives from and is based on the natural science model. Over time it has emerged and been legitimated as the only path to cumulative scientific knowledge. While specialists in evaluation research may vary in their emphasis on cost-benefit analysis, experimental design, multiple regression analysis, the construction of simulated mathematical models, systems analysis, survey research, standardized measurement, and input-output analysis, the underlying paradigm--*The* Scientific Me-

Michael Patton is a postdoctoral fellow in evaluation research methodology and assistant professor of sociology at the University of Minnesota.

thod--remains basically the same for all of these techniques. Before pursuing the main concern of this paper, it may be helpful to briefly outline that Method as operationalized by some of its major advocates in evaluation research and suggest some reasons for its dominance.

1

The Dominant Paradigm

A useful case in point is a study by Bernstein and Freeman (1974), a mammoth evaluation of evaluation research, sponsored and published by the Russell Sage Foundation. Their focus included all evaluation studies directly funded by agencies of the federal government in the fiscal year of 1970. They sampled the population of all large-scale social-action programs aimed at ameliorating some social problem in the areas of health, education, welfare, public safety (crime), income security, housing, and manpower and carrying a minimum research budget of $10,000. Their final analysis was based on 236 evaluation research projects.

TABLE I. BERNSTEIN AND FREEMAN (1974) CODINGS OF EVALUATION QUALITY VARIABLES

Variable Measuring Some Aspect of Evaluation Quality	*Coding Scheme (where higher coding number represents higher quality)*
1. Nature of Research Design	0 = Descriptive Study
	1 = Comparative, longitudinal or cross-sectional studies without randomization or control
	2 = Experimental designs without both randomization or control
	3 = Experimental designs with randomization and control
2. Representativeness of the Sample	0 = Haphazardly drawn samples
	1 = Moderately representative

3.	Sampling	0 = Non-systematic, non-random, non-systematic random, and random or non-random cluster samples
		1 = Stratified random, simple random, or all (i.e. universe)
4.	Type of Data Analysis	0 = No statistics, ratings, or impressions
		1 = Narratives or impressionistic summaries
		2 = Rating from qualitative data
		3 = Simple descriptive statistics
		4 = Multivariate statistics
5.	Nature of Data Analysis*	0 = Qualitative Analyses
		1 = Evenly divided between qualitative and quantitative analyses
		2 = Quantitative analyses
6.	Quality of Measurement Procedures**	0 = Inadequate measurement
		1 = Adequate measurement

---

*Explanatory Quote from Bernstein and Freeman:

While there may be some debate as to the order we have imposed here, i.e. quantitative as higher than half quantitative and half qualitative, we feel justified in so doing since most of the current literature on evaluation research methods, e.g., Suchman, E.A., *Evaluative Research*, 1967, Russell Sage, N.Y.; Caro F. (ed.), *Readings in Evaluation Research*, 1971 Russell Sage, N.Y.; Rossi, P. and Williams W., *Evaluating Social Programs*, 1972, Seminar Press, N.Y.; and Sheldon, E.B. and Bernstein, I.N., "Methods of Evaluative Research", in *Social Science Methods*, (ed.) Robert Smith, 1973, Free Press, N.Y., strongly suggests that the best evaluations in terms of research quality are those which are highly quantitative.

In reviewing the findings of the Bernstein-Freeman study, which set out to assess the *quality* of evaluation research projects, what is of immediate interest to us is the way the study measured 'quality': the quality variables they identified and measured represent a fully explicit description of the dominant evaluation research paradigm. Table I shows how they coded their six major indicators of quality, with a *higher* code number representing *higher* quality evaluation research. What emerges is 1) experimental designs with randomization and control groups, 2) reliable and valid measurement instrumentation, 3) representative samples that are 4) randomly selected, and 5) sophisticated statistical analysis of 6) completely quantitative data.

Some might want to add to the Bernstein-Freeman list but few practicing social scientists would question the accuracy or validity of the research paradigm they describe as *The* Scientific Method. (Some may note with dismay the absence of any measurement to indicate whether the information collected was relevant to the programs evaluated, whether the evaluation information was used by decision-makers and program participants, whether the outcomes measured were those held to be important by program funders, administrators, and participants, or whether the evaluation design and results were understandable to those for whom the evaluation was conducted. The Bernstein-Freeman paradigm, however, makes no pretense of addressing such questions; in the dominant paradigm such questions are not of central methodological interest, a point to which I return later.) At a conference on evaluation and policy research sponsored by the American Academy of Arts and Sciences in 1969, Peter Rossi reported general consensus about the most desired evaluation research methods. The consensus was virtually identical to the model found most desirable by Bernstein and Freeman. A cursory skimming of major educational and social science research journals yields a similar lack of disagreement. In their widely used methodological primer, Campbell and Stanley (1966:3)

**Explanatory footnote from Bernstein and Freeman:

The satisfaction of a measure having adequate content validity as it appears in Kerlinger, Fred, *Foundations of Behavioral Research*, 1964, N.Y.: Holt, Rinehart and Winston, pp. 444-447. An example of response which was coded adequate was: "The criteria by which the effectiveness of an educational program aimed at increasing cognitive ability of mentally retarded children was the use of standardized reading comprehension, vocabulary, and arithmetic tests, all of which had been pretested for reliability on other similar target populations. Five repeated measures were taken over a two-year period."

call this paradigm "the only available route to cumulative progress."

What accounts for such certainty, or at any rate such acceptance, of an intellectual construct, at a time when natural scientists themselves are reexamining their most fundamental propositions? The answer to that question may not be so inaccessible. As Kuhn (1970:80) explains, "A paradigm governs, in the first instance, not a subject matter but rather a group of practitioners." Those practitioners most committed to the dominant paradigm are found in the universities where they not only employ *The* Scientific Method in their own evaluation research, but where they also nurture students in a commitment to that same methodology (cf. Bernstein and Freeman, 1974).

There are other reasons for the dominance of the natural science model, reasons that go somewhat beyond the merits of the Method. William J. Filstead (1970:3-4) suggests such reasons as "ego fulfillment; the achievement of scientific respectability; the quest for social status on a par with that of natural scientists; and grantsmanship, which, although it is not necessarily helpful in ascertaining the validity of the data, does enhance both those who collect data in the appropriate fashion and the discipline that fosters adherence to those appropriate methods of data collection."

While there can be some argument about the reasons for the dominance of the natural science model in educational social scientific research, the fact of the dominance cannot be seriously doubted. The issue for us is that *the very dominance of The Scientific Method in evaluation research appears to have cut off the great majority of its practitioners from serious consideration of any alternative research paradigm.* The label 'research' has come to mean the equivalent of employing *The* Scientific Method--of working within the dominant paradigm.

2

The Alternative Paradigm

The discussion that follows is focused on broad epistemological contrasts.* The alternative (methodological) paradigm that I shall discuss is drawn together from a number of emerging directions--trends, ideas, approaches, methods, and perspectives that are not always clearly articulated by their adherents. It draws on work in qualitative methodology, phenomenology, symbolic interactionism, *Gestalt* psychology, ethnomethodology, and the general notion or doctrine of *verstehen*. Kenneth Strike (1972:28) describes this tradition as follows:

> The basic dispute clustering around the notion of *verstehen* has typically sounded something like the following: The advocate of some version of the *verstehen* doctrine will claim that human beings can be understood in a manner that other objects of study cannot. Men have purposes and emotions, they make plans, construct cultures, and hold certain values, and their behavior is influenced by such values, plans, and purposes. In short, a human being lives in a world which has "meaning" to him, and, because his behavior has meaning, human actions are intelligible in ways that the behavior of nonhuman objects is not. The opponents of this view, on the other hand, will maintain that human behavior is to be explained in the same manner as is the behavior of other objects of nature. There are laws governing human behavior. An action is explained when it can be subsumed under some such law, and, of course, such laws are confirmed by empirical evidence.

The alternative paradigm stresses understanding that focuses on the *meaning* of human behavior, the context of social interaction, an *emphatic* understanding of subjective (mental, not nonobjective) states, and the connection between subjective states and behavior. Filstead explains that the tradition of *verstehen* or understanding "has had its greatest influence in formulating the position that recognizes the importance of both an inner and an outer perspective of human behavior....The inner perspective places emphasis on man's ability to know himself and, hence, to know and understand others through 'sympa-

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

thetic introspection," and 'imaginative reconstruction' of 'definitions of the situation.'"

The alternative paradigm proposes an active, involved role for the social scientist/evaluation researcher. "Hence, insight may be regarded as the core of social knowledge. It is arrived at by being on the inside of the phenomena to be observed....It is participation in an activity that generates interest, purpose, point of view, value, meaning, and intelligibility, as well as bias" (Wirth, 1949:xxii). As Filstead (1970:4) says, "this in no way suggests that the researcher lacks the ability to be scientific while collecting the data. On the contrary, it merely specifies that it is crucial for validity--and, consequently, for reliability--to try to picture the empirical social world as it actually exists to those under investigation, rather than as the researcher imagines it to be." More concretely, the alternative paradigm relies on field techniques from an anthropological rather than natural science tradition, techniques such as participant observation, in-depth interviewing, detailed description, and qualitative field notes.

3

Opposing Paradigms

I have now described the broad outlines of two contrasting evaluative research paradigms. It is the task of the remainder of this paper to sharpen these contrasts, to bring them into high relief, *to make them appear as opposites.* Such an analysis, based on non-existing ideal-types, will clearly overstate the case. Tacit understandings about flexible parameters will here appear as absolute rules of procedures. Areas of mutuality, common concern, and similarity of commitments will be largely ignored.

The justification for such an approach can be found in the very nature of paradigms. A paradigm is a world view, a general perspective, a way of breaking down the complexity of the real world. As such, paradigms are deeply embedded in the socialization of adherents and practitioners telling them what is important, what is legitimate, what is reasonable. Paradigms are normative, they tell the practitioner what to do without the necessity of long existential or epistemological consideration. But it is this aspect of a paradigm that constitutes both its strength *and* its weakness--its strength in that it makes action possible, its weakness in that the very reason for action is hidden in the unquestioned assumptions of the paradigm. It is to raise these assumptions to the level of consciousness among evaluation researchers that this analysis is undertaken. The difficulty of this task is clear from Kuhn's description of the power of paradigms:

> Scientists work from models acquired through education and through subsequent exposure to the literature often without quite knowing or needing to know what characteristics have given these models the status of community paradigms. And because they do so, they need no full set of rules. The coherence displayed by the research tradition in which they participate may not imply even the existence of an underlying body of rules and assumptions that additional historical or philosophical investigation might uncover. *That scientists do not usually ask or debate what makes a particular problem or solution legitimate tempts us to suppose that, at least intuitively, they know the answer. But it may only indicate that neither the question nor the answers are felt to be relevant to their research. Paradigms may be prior to, more binding, and more complete*

than any set of rules for research that could be unequivocally abstracted from them. (Kuhn, 1970:46.)

It is because "paradigms may be prior to, more binding, and more complete than any set of rules for research that can be unequivocally abstracted from them" that the analysis here will focus on dominant motifs, modalities of thought and action, and illumination of tacit understandings. The dichotomies constructed will be aimed at capturing the underlying and fundamental elements in the two paradigms which are the bases of their opposition and competition.

At the outset I considered the possibility of attempting to describe and contrast the two paradigms in a neutral fashion. However, the very dominance of one paradigm, the natural science model, and the subordination of the second paradigm, the alternative paradigm, convinced me that it is more important to attack this imbalance than to maintain neutrality. My concern here is two-fold: First, I am concerned that practitioners and adherents of the dominant paradigm show little awareness of or consciousness about even the existence of an alternative paradigm; and secondly, I am concerned that practitioners of the dominant paradigm seem to be insensitive to and unaware of the degree to which their methodology is based upon a relatively narrow philosophical/ideological/epistemological view of the world. "It is important," Mills wrote, "to get this point quite clear, for one would suppose that philosophical tenets would not be central to the shaping of an enterprise which is so emphatic in its claim to be Science. It is important also because the practitioners of the style do not usually seem aware that it is a philosophy upon which they stand" (Mills, 1961:56).

It is in this context that I wish to approach the following discussion of evaluative research paradigms. The assets of the alternative paradigm need to be stressed and the shortcomings of the dominant paradigm need to be seriously examined for the majority of evaluation researchers seem to be oblivious of the assets of the former, and euphoric about the techniques of the latter. Hubert Blumer (1969:47) put the issue this way: "This opposition needs to be stressed in the hope of releasing social scientists from unwitting captivity to a format of inquiry that is taken for granted as the naturally proper way in which to conduct scientific inquiry...."

As a final introductory note I would add that there is a tension in this analysis between the abstract and the concrete. I have tried to overcome it by illustrating the points of paradigm opposition with examples from the literature on educational evaluation, drawing particularly on evaluations that have to do with open education and other alternatives to traditional schooling. These examples help make a point that runs throughout this analysis: If there can be clarity about the need to adopt evaluation methods that suit the nature of the program being evaluated, the contrasting natures of

traditional and open education programs suggest a need for contrasting evaluation strategies and techniques. I shall pursue this point throughout the analysis that follows.

4

Qualitative vs. Quantitative Methodology

Kuhn (1970:184-5) in his discussion of science paradigms notes that the values held by scientists function to help them choose between incompatible ways of practicing their discipline and that "the most deeply held values concern predictions: they should be accurate; quantitative predictions are preferable to qualitative ones...." Kuhn is writing mainly about natural scientists, but it is clear that the values of natural scientists concerning prediction have been enthusiastically embraced by social scientists and educational researchers: Not only are quantitative predictions preferable to qualitative ones, but qualitative analyses in general have little legitimacy beyond certain limited exploratory situations.

The art and science of quantification constitutes the very core of the dominant paradigm. To turn words into numbers, historical trends into prediction equations, and the behavior of people into probability tables and standardized regression coefficients--these are the greatest miracles in Science, and to the performers of these miracles go the greatest of all Scientific rewards: recognition and high status.

The methodological status hierarchy in Science is clear: the harder the data, the more scientific the results and the higher the status. (By "hardness of data" is meant the degree to which you can assign numbers to what you are studying and manipulate those numbers using sophisticated statistical techniques.) Thus, economics outranks sociology as Science. Within sociology the demographers, empiricists, and quantitative methodologists rank at the top of the methodological status hierarchy; ethnomethodologists, participant observers, and qualitative methodologists occupy the lower parts of the hierarchy, meaning they have more difficulty getting their work published, greater problems on the job market, less agility at attaining tenure and promotion, and greater difficulty obtaining research grants. The same methodological status hierarchy rules other disciplines, including Schools of Education.

The foregoing is not meant as an across-the-board attack on the use of statistics in evaluation research. *The problem is the use of statistics to the virtual exclusion of other types of data.* In this regard, C. Wright Mills (1961:50) observed that the dominance of statistical methodology has led to a "methodological in-

hibition" that he called "abstracted empiricism." The problem with abstracted empiricism is that "it seizes upon one juncture in the process of work and allows it to dominate the mind."

The dominance of quantitative methodology has acted to severely limit the kinds of questions that are asked* and the types of problems that are studied. While most phenomena are not necessarily intrinsically impossible to measure quantitatively, certain types of phenomena are clearly easier to measure numerically than others. It is easier, for example, to measure the number of words that a child spells correctly than to measure that same child's ability to use those words in a meaningful way. The vast majority of educational researchers have clearly opted for the first procedure. It is easier to count the number of minutes a student spends reading books in class than it is to measure what reading *means* to that child. We have a large number of studies of the former, but we know little about the latter.

Quantitative methodology assumed the possibility, desirability, and even the necessity of applying some underlying empirical standard to social phenomenon. Thus, an underlying standard of measurement can be applied to measure the wavelength of blue light. But qualitative methodology assumes that some phenomena are not amenable to such mediation. While you can measure the length of blue light, can you capture in quantitative notation what the color blue looks and feels like? The experience of looking at blue light is a direct encounter between phenomenon and observer; it is not easily amenable to statistical measurement.

The point here is that different kinds of problems require different types of research methodology. If all we want to know is the number of words a child can spell or the frequency of interaction between children of different races in desegregated schools, then statistical procedures are appropriate. However, if we want to understand the relevance of the words to that child's particular life or the meaning of inter-racial interactions then some form of qualitative methodology (participant observation, in-depth interviewing, systematic field work) which allows the researcher to obtain firsthand knowledge about the empirical social world in question may well be more appropriate. Mills (1961:73-74) has stated this approach quite succinctly:

> If the problems upon which one is at work are readily amenable to statistical procedures, one should always try to use them....No one, however, need accept such procedures, when generalized, as the only procedure available. Certainly no one need accept this model as a total canon. It is not the only empirical manner.
>
> It is a choice made according to the requirements of our problems, not a 'necessity' that follows from an epistomological dogma.

*See also Hein, G.E., *An Open Education Perspective on Evaluation*, in this series.

An extended example may help illustrate the importance of seeking congruence between the phenomenon studied and the research methodology employed for this study. The example, a major study frequently quoted, concerns the key issue of whether or not educational innovation makes a difference in children's achievement. After examining some four decades of educational research, John Stephens (1967) concluded that educational innovation makes little difference. "But," asks Edna Shapiro (1973:542), "can such a judgment be made when the researcher has sampled only an extremely narrow band of measurement within a constant and equally restrictive situation?"

Shapiro asked this question after finding no differences in achievement test scores between 1) children in an enriched Follow Through (FT) program modeled along the lines of open education and 2) children in comparison schools not involved in Follow Through or other enrichment programs. *When the children's responses in the test situation were compared, no differences of any consequence were found. However, when observations were made of the children in their classrooms, there were striking differences between the Follow Through and comparison classes:*

> A satisfactory explanation of the outcomes of this study raises general questions about assessing the impact of educational programs. Other studies may be more elaborately mounted, more carefully controlled, more elegantly analyzed, but the basic issues remain the same. In this study, when we observed the children in their classrooms, there were striking differences between the FT and comparison classes; when we compared the children's responses in the test situation, there were no differences of any consequence. Conventional explanations would make little of the classroom differences, stressing the absence of difference in individual test response. The conventional explanation for equivocal findings (and they are not unique--the educational research literature is replete with negative findings) is that the programs being compared do not make a difference, that the research design was inadequate, or that it is naive to expect differences since program variations do not make a noticeable difference. My contention is that such explanations do not go far *enough. While it is important to try to explain negative test results, it is far more important to account for the disparity between the negative test finds and the clear differences observed in classroom behavior.* (Shapiro, !973:527.)

Based on systematic observations the Follow Through classrooms "were characterized as lively, vibrant, with a diversity of curricular projects and children's products, and an atmosphere of friendly, cooperative endeavor. The non-FT classrooms were characterized as relatively uneventful, with a narrow range of curriculum, uniform ac-

tivity, a great deal of seat work, and less equipment; teachers as well as children were quieter and more concerned with maintaining or submitting to discipline." (Shapiro, p. 529.) Observations also revealed that the children behaved differently in these two types of environments. Yet standardized achievement tests failed to detect these differences. Shapiro (p. 532) suggests that "there were factors operating *against* the demonstration of differences," which call into question traditional ways of gauging the impact and effectiveness of different kinds of school experience. *The testing methodology, in fact, narrowed the nature of the questions that were being asked and pre-determined non-significant statistical results.* Shapiro's analysis of how the quantitative methodological procedures determined the research results are so insightful and so important that we quote her at length:

> Studies of the effectiveness of different kinds of educational programs share a common methodology: children of comparable background and ability are exposed to or participate in experiences which vary in certain ways and are subsequently tested on aspects of learning or performance presumed to demonstrate the impact of the differences in their experiences....
>
> In this study, too, the child's responses in the test situation were considered critical. What children do in the classroom--the kinds of questions they ask, the kinds of activities they engage in, the kinds of stories, drawings, poems, structures they produce, the kinds of relationships they develop with other children and the teacher--indicates not only what they are capable of doing but what they are allowed to do. Classroom data are generally down-graded in attempts to study the effects of educational programs because we cannot know whether the comparison group, given the same opportunities, would behave in similar ways. And conversely, we do not know whether, if the opportunity were removed, there would be any carry-over to a new classroom situation, that is, whether the effects have been internalized. Nor is it easy to separate the contribution of and effect upon individual children in the group. Following the line of reasoning of an earlier study, I assumed that the internalized effects of different kinds of school experience could be observed and inferred only from responses in test situations, and that the observation of teaching and learning in the classroom should be considered auxiliary information, useful chiefly to document the differences in the children's group learning experiences.
>
> The rationale of the test, on the contrary, is that each child is removed from the classroom and treated equivalently, and differences in response

are presumed to indicate differences in what has been taken in, made one's own, that survives the shift to a different situation.

The findings of this study, with the marked disparity between classroom responses and test responses, have led me to reevaluate this rationale. This requires reconsideration of the role of classroom data, individual test situation data, and the relation between them. *If we minimize the importance of the child's behavior in the classroom because it is influenced by situational variables, do we not have to apply the same logic to the child's responses in the test situation, which is also influenced by situational variables?*

The individual's responses in the test situation have conventionally been considered the primary means to truth about psychological functioning. Test behavior, whether considered as a sign or sample of underlying function, is treated as a pure measure. Yet the test situation is a unique interpersonal context in which what is permitted and encouraged, acceptable and unacceptable, is carefully defined, explicitly and implicitly. *Responses to tests are therefore made under very special circumstances. The variables that influence the outcome are different from those which operate in the classroom, but the notion that the standard test or interview provides equal treatment for all subjects is certainly open to question.* (Shapiro, pp. 532-534.)

Shapiro elaborates and illustrates these points at considerable length. Her conclusion goes to the heart of the problem posed by the dominance of a single methodological paradigm in evaluation research: *"Research methodology must be suited to the particular characteristics of the situations under study....An omnibus strategy will not work"* (p. 543, italics added).

Most social scientists do not deny the immense heuristic value of qualitative data. What they do deny is that qualitative methodology can be a legitimate source of either data collection, systematic evaluation, or theory construction. At best, social scientists are willing to recognize that qualitative methodology may be useful at an exploratory stage of research prefatory to quantitative research. However, "to force all of the empirical world to fit a scheme that has been devised for a given segment of that world is philosophical doctrinizing and does not represent the approach of a genuine empirical science." (Blumer, 1969:23).

There is indeed a viable alternative to the dominant natural science model, an alternative that not only employs different methods but also asks different questions. And, as Kuhn has explained, one of the functions of scientific paradigms is to provide criteria for choosing problems that can be assumed to have solutions: "Change in the standards governing permissible problems, concepts, and explanations

can transform a science" (p. 106). It is the failure of the dominant natural science paradigm to answer important questions like those raised by Shapiro that makes serious consideration of the alternative paradigm so crucial for evaluation research.

5

Reliability vs. Validity

Any consideration of paradigms in science must focus on dominant motifs and patterns. Paradigms tell scientists what to emphasize, what to look for, what questions to be concerned with, and what standards to apply. *Competing paradigms raise questions of emphasis.* It is the contention of this paper that the dominant paradigm in scientific research, with its quantitative emphasis, has been preoccupied with reliability, while the alternative paradigm emphasizes validity.

Reliability concerns the replicability and consistency of scientific findings. One is particularly concerned here with inter-rater, inter-item, interviewer, observer, and instrument reliability. Validity, on the other hand, concerns the meaning and meaningfulness of the data collected and instrumentation employed. Does the instrument measure what it purports to measure? Does the data mean what we think it means?

Merton (1957:448), one of the most prominent godfathers of sociology, argues that the cumulative nature of science requires a high degree of consensus among scientists and leads, therefore, to an inevitable enchantment with problems of reliability. With the proposition that scientific research has been preoccupied with questions of reliability, I can agree; but I part company with the proposition that such a preoccupation is necessary and good.

Irwin Deutscher (1970:33) has stated the problem with great cogency:

> We have been absorbed in measuring the amount of error which results from inconsistency among interviewers or inconsistency among items on our instruments. We concentrate on consistency without much concern with what it is we are being consistent about or whether we are consistently right or wrong. As a consequence we may have been learning a great deal about how to pursue an incorrect course with a maximum of precision.
>
> It is not my intent to disparage the importance of reliability per se; it is the obsession with it to which I refer. Certainly zero reliability must result in zero validity. But the relationship is not linear, since infinite perfection of reliability (zero error) may also be associated with zero

> validity. Whether or not one wishes to emulate the scientist and whatever methods may be applied to the quest for knowledge, we must make our estimates of, allowances for, and attempts to reduce the extent to which our methods distort our findings.

The problem with the standardized tests in Shapiro's study of open education Follow Through classrooms was not that they were unreliable, but that they were not valid measures of the learning taking place in those classrooms. Yet any suggestion that standardized tests may be an inappropriate measure of learning is met with outraged accusations that reliability of measurement is being sacrificed.

At the same time, validity has become a function of frequency of use of some instrument. The often-used and highly reliable instrument takes on a sanctity that places it above question. After a while we lose sight of the actual behaviors that are supposed to be associated with the instrument. "The widespread misconceptions about the so-called IQ provide a particularly flagrant example of such a dissociation. One still hears the term 'IQ' used as though it referred, not to a test score, but to a property of the organism" (Anastasi, 1973:xi).

When one actually looks at the operational definitions and measures of major educational and social scientific concepts, one sees that their transparency and bias are frequently astounding though their reliability is extremely high. In addition, we seem to have lost sight of the fact that responses *mean* different things in different settings and different contexts. (The only way to discern such variations in shades of meaning is to directly interact with and observe respondents in various relevant settings.) Thus, instruments prepared for evaluation in one setting are adopted for evaluation in other settings with a facility that shows arrogant insensitivity to the issue of cross-setting validity. This does not mean that *every* evaluation must include development of new instrumentation. But *every* evaluation must include some effort to establish the validity of the instrumentation adopted for the setting in which it is used.

The alternative evaluation paradigm makes the issue of validity central by getting close to the data, being sensitive to qualitative distinctions, attempting to develop empathy with program participants and thereby approaching the data subjectively, and taking a holistic and process perspective on evaluation (issues taken up later in this paper). The overriding issue in the *verstehen* approach to science is the *meaning* of the scientist's observations and data, particularly its meaning for participants themselves. The constant focus is on a valid representation of what is happening, not at the expense of reliable measurement, but without allowing reliability to determine the nature of the data.

Discussion of varying emphasis on reliability and validity in the two paradigms is particularly difficult because the ideal in both paradigms is high reliability

and high validity. Nevertheless, differences in emphasis in the two paradigms are clearly discernible. The differences are a matter of emphasis and attention, but it is of such differences that alternative paradigms are made.

6

Objectivity vs. Subjectivity

Objectivity is considered the *sine qua non* of *The* Scientific Method. Qualitative methodology and a phenomenological approach to evaluation research, on the other hand, most frequently stimulate charges of subjectivity--a label held to be the very antithesis of scientific inquiry. To be subjective means to be biased, unreliable, and non-rational. Subjective data imply opinion rather than fact, intuition rather than logic, impression rather than confirmation. Social scientists are encouraged to eschew subjectivity in favor of making their work "objective and value-free."

Some evaluation researchers recognize that social action research may take one so close to questions of politics and values that it may be impossible to completely eliminate subjectivity. Under these conditions "the task for the development of evaluative research as a 'scientific' process is to 'control' this intrinsic subjectivity, since it cannot be eliminated..., to examine the principles and procedures that man has developed for controlling subjectivity--the scientific method...." (Suchman, 1967: 11-12).

Not surprisingly, the means for controlling subjectivity through the scientific method are the techniques of the dominant paradigm, particularly quantitative methodology and emphasis on reliability. Yet we have already argued that quantitative methodology works, in practice, to limit and even bias the kinds of questions that can be asked and the nature of admissible solutions. In effect, identifying objectivity as the major virtue of the dominant paradigm is an ideological statement the function of which is to legitimize, preserve, and protect the dominance of a single evaluation methodology.

Michael Scriven (1972:94) asserts that quantitative methods are no more synonymous with objectivity than qualitative methods are synonymous with subjectivity. "Errors like this are too simple to be explicit. They are inferred confusions in the ideological foundations of research, its interpretations, its applications." Scriven goes on to comment that "it is increasingly clear that the influence of ideology on methodology and of the latter on the training and behavior of researchers and on the identification and disbursement of support is staggeringly powerful. Ideology is to research what Marx suggested the

economic factor was to politics and what Freud took sex to be for psychology."

Scriven's (1972) discussion of "Objectivity and Subjectivity in Educational Research" is a major contribution in the struggle to detach the notions of objectivity and subjectivity from their traditionally narrow associations with quantitative and qualitative methodology, respectively. He presents a cogent argument for recognizing as legitimate science not only the *prediction* of social phenomena but also, and perhaps even more important, recognizing as science the pursuit of understanding--*verstehen*. The quest for social prediction in the same sense as prediction operates in the classical natural science paradigm is "pipe dreaming" (p. 115). The practice of Science has led to a formalistic split between the mental and the logical, seen as the subjective and the objective, which keeps researchers from seeing that "*understanding*, properly conceived, is in fact an 'objective' state of mind or brain and can be tested quite objectively; and it is a functional and crucial state of mind, betokening the prescence of skills and states that are necessary for survival in the sea of information. There is nothing wrong with saying, in this case, that we have simply developed an enlightened form of intersubjectivism. But one might also equally well say that we have developed *an enlightened form of subjectivism--put flesh on the bones of empathy*" (p. 127).

Scriven is here suggesting two different ways of looking at the same thing. The idea of dual perspectives concerning a single phenomenon goes to the very heart of the dichotomy between paradigms. Two scientists may look at the same thing, but because of different theoretical perspectives, different assumptions, or different ideology-based methodologies, they may literally *not* see the same thing (cf. Petrie, 1972: 8). Indeed, Kuhn (1970: 113) argues that "something like a paradigm is prerequisite to perception itself. What a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see. In the absence of such training there can only be, in William James' phrase, 'a bloomin' buggin' confusion.'"

It is in this context that the dominant paradigm's assertion of objectivity can be called ideology. Such an analysis is based on the relativistic assumption that it is not possible for us to view the complexities of the real world without somehow filtering and simplifying those complexities. That act of filtering and simplifying affects what the observer sees because it necessarily brings into play the observer's past experiences of the world. In the final analysis, this position means that we are always dealing with perceptions, not 'facts' in some absolute sense. As Petrie (1972:49) put it, "the very categories of things which comprise the 'facts' are theory dependent" or, in our terms, paradigm dependent. It is this recognition that the scientist inevitably operates within the constraints of a perception-based paradigm

(with ideological and political underpinnings) that leads Howard Becker (1970:15) to argue that "the question is not whether we should take sides, since we inevitably will, but rather whose side we are on."

It is also in this context that the notion of subjectivity, properly construed, can become a positive rather than a pejorative term in evaluation research. Subjectivity in the alternative paradigm "allows the researcher to 'get close to the data,' thereby developing the analytical, conceptual, and categorical components from the data itself--rather than from the preconceived, rigidly structured, and highly quantified techniques that pigeonhole the empirical social world into the operational definitions that the researcher has constructed" (Filstead, 1970:6). Moreover, *a positive view of subjectivity--getting close to and involved with the data--makes it possible for evaluation researchers to take into account their personal insights and behavior.* As Scriven (1972:99) laments, "For the social sciences to refuse to treat their own behavior as data from which one can learn is really tragic." Alvin Gouldner (1970) is even more adamant on this point. He suggests that "high science methodology" creates a gap between what the researcher as scientist deals with and what that same researcher (like others) confronts as an ordinary person, experiencing his or her *own existence:*

> It is a function of high science methodologies to widen the gap between what the sociologist is studying and his own personal reality. Even if one were to assume that this serves to fortify objectivity and reduce bias, it seems likely that it has been bought at the price of the dimming of the sociologist's self-awareness. In other words, it seems that at some point, the formula is: the more rigorous the methodology, the more dimwitted the sociologist; the more reliable his information about the social world, the less insightful his knowledge about himself (p. 56).

To say that the evaluation researcher can learn much by getting close to the data is not to say that there is no systematic way of conducting scientific inquiry, that anything goes. The point, rather, is to bring the mind and feelings of the human being back into the center of evaluation research--a center that has thus far been dominated by techniques and rules. It is to recognize that science is really nothing if it is not the application of critical intelligence to critical problems. The narrow parameters of the dominant paradigm have constrained that critical intelligence under the guise of attaining a natural science objectivity. In this regard, C. Wright Mills (1961:58) quotes Nobel Prize-winning physicist Percy Bridgman to the effect that "there is no scientific method as such, but the vital feature of the scientist's procedure has been merely to do his utmost with his mind, *no holds barred.*"

The *verstehen* or understanding approach to scientific inquiry is based on the application of critical intelligence to social phenomena without mediation by preconceived categories and without the abstraction of numerical representation. This alternative paradigm seeks to redraw the boundaries of legitimate scientific inquiry thereby increasing the domain of what has been labeled (qualitatively) subjective by the dominant paradigm so that many of what have been thought of as illegitimate practices and topics can be tackled.

Space does not permit a full epistemological exploration of the arguments underlying traditional notions of objectivity and subjectivity in evaluation research. It may be helpful, however, to again use the problem of evaluating innovations in open education to illustrate the different perspectives on objectivity and subjectivity represented by the two evaluation methodology paradigms. The dominant paradigm lauds the use of standardized tests to measure pupil achievement in school because these tests are highly reliable, their outcomes have been widely replicated on varying populations, and their statistical properties are well-known. In brief, standardized tests represent an objective measure of achievement across situations and populations. Standardized tests properly administered minimize the introduction or researcher bias in measuring achievement.

However, standardized tests can bias evaluation results by imposing a standardized and controlled stimulus in an environment where learning depends on spontaneity, creativity, and freedom of expression, as Shapiro (1973) found in her study of innovative Follow Through classrooms described earlier. Moreover, she found that the results of the test measured response to a stimulus (the test) which was essentially alien to the experience of the children. Because the open classroom relies substantially less on paper-and-pencil skills and because student progress is monitored on a personal basis without the use of written examinations, student outcomes in the open classroom could not be "objectively" measured by standardized tests. Such tests fail to delineate the learning outcomes of children who make differential uses of particular classroom situations. Shapiro argues that "the quest for objective control over the multiplicity of interdependent events occurring in a classroom has led to a concentration on ever smaller units of behavior, divorced from context and sampled in rigorously scheduled time units (p. 543)."

The actual behaviors of children observed in the open classroom situation were not validly captured by standardized tests or one-to-one interviews with adults, even when the interviewer was someone who was familiar to the children. For the children in open classrooms, "the transition from the relatively free and easy exchange of the classroom to the more constricted interview was not automatic; it was, in fact, not possible (p. 539)." Del Hymes (1971:56) describes this kind of situation in more techni-

cal language: "When a child from one developmental matrix enters a situation in which the communicative expectations are defined in terms of another, misperception and misanalysis may occur at every level...; intents and innate abilities may be misevaluated because of differences of systems for the use of language and for the import of its use (as against other modalities)."

The problem is not simply one of finding a new or better standardized test. The problem is one of understanding the context of observed behaviors, the meaning of specific achievement outcomes to the child in a more holistic setting than is possible with any standardized test. This does not mean that standardized tests may not be useful for certain specific questions, but they are not sufficient when the issue is *understanding*, not just prediction. *Understanding in its broadest sense requires getting close enough to the situation to gain insight into mental states; it means subjectivity in the best scientific sense of the term.* The alternative paradigm seeks to legitimize and incorporate this subjectivity into evaluation research, not to the exclusion of the methodology of the dominant paradigm, but in addition to it.

If a limited notion of subjectivity based on careful and systematic observation by trained researchers in the best tradition of anthropological research cannot be made a legitimate part of evaluation research, then a host of crucial questions will be excluded from investigation. "If we cannot straighten out the situation," Scriven (1972:97) warns, "we are doomed to suffer from the swing of the pendulum in the other direction, a swing which it is easy to see implicit in the turn toward irrationalistic, mystical, and emotional movements thriving in or on the fringes of psychology today. There is much good in them on their own merits, but the ideology that is used to support them is likely to breed the same intolerance and repression that the positivists spread through epistemology and psychology for a quarter century."

7

Distance from vs. Closeness to the Data

There are several additional paradigm components that have emerged in the discussion of quantitative versus qualitative methodology, and reliability versus validity, and objectivity versus subjectivity that deserve additional comment. One of these involves the issue of how close the investigator should get to the data. The dominant paradigm prescribes distance to guarantee neutrality and objectivity. This component of the dominant paradigm has become increasingly important with the professionalization of the social sciences and educational research establishment. Professional comportment connotes cool, calm, and detached analysis without personal involvement. The profession is identified by and takes pride in its skills--in this case quantitative methodology and empiricism--not in its ability to serve the needs of 'clients' (cf. Horowitz, 1964:10-11).

Alvin Gouldner (1970:53) suggests that this emphasis on detachment and professional distance is the social scientist's way of accommodating himself to his alienation in contemporary society, a reaction to "man's failure to possess the social world that he created." This alienation is built on the notion that society and culture can be viewed like any other 'natural' phenomena, as having laws that operate quite apart from the intentions, motivations, and plans of human beings. Methodology follows this assumption by emphasizing prediction and universal laws rather than understanding and human meaning. Horowitz (1965:11) is less kind, emphasizing the elitism and arrogance of social scientists as they disguise their search for status and professional prestige behind a thin veil of neutrality and detachment.

Whatever the source of the emphasis on distance and detachment in the dominant paradigm, its centrality to that methodology can scarcely be questioned. What is questioned by the alternative paradigm is the necessity of distance and detachment. The alternative paradigm assumes that without empathy and sympathetic introspection derived from personal encounters the observer cannot fully understand human behavior. Understanding comes from trying to put oneself in the other person's shoes, from trying to discern how others think, act, and feel. John Lofland (1971) explains that methodologically this means 1) getting close to the people being studied through attention to the minutia of daily life, through physical

proximity over a period of time, and through development of closeness in the social sense of intimacy and confidentiality; 2) being truthful and factual about what is observed; 3) emphasizing a significant amount of pure description of action, people, activities, etc.; and 4) including as data direct quotations from participants as they speak and/or from whatever they might write. "The commitment to get close, to be factual, descriptive, and quotive, constitutes a significant commitment to represent the participants *in their own terms*" (p. 4).

The commitment to closeness is further based upon the assumption that the inner states of people are important and knowable. It is at this point that the alternative paradigm intersects with the phenomenological tradition (cf. Bussis, Chittenden, and Amarel, 1973). Attention to inner perspectives does not mean administering attitude surveys. "The inner perspective assumes that understanding can only be achieved by actively participating in the life of the observed and gaining insight by means of introspection" (Bruyn, 1963:226).

A commitment to get close to the data and a willingness to capture participants in their own terms implies an openness to the phenomenon under study that is relatively uncontaminated by preconceived notions and categories. "In order to capture the participants 'in their own terms' one must learn *their* analytic ordering of the world, *their* categories for rendering explicable and coherent the flux of raw reality. That, indeed,is the first principle of qualitative analysis" (Lofland, 1971:7).

In the Shapiro study of open Follow Through classrooms, it was her closeness to the classrooms under study and the children in those classrooms that allowed her to see that something was happening that was not captured by standardized tests. She could *see* differences in children. She could *understand* differences in the meaning of their different situations. She could feel their tension in the testing situation and their spontaneity in the more natural classroom setting. Had she worked solely with data collected by others, had she worked only at a distance, she would never have discovered the crucial differences in the classroom settings she studied--differences in modes of achievement which actually allowed her to *evaluate* the innovative program in a meaningful and relevant way.

Again, it is important to note that the admonition to get close to the data is in no way meant to deny the usefulness of quantitative methodology. Rather, it is to say that statistical portrayals must always be interpreted and given human meaning. That many quantitative methodologists fail to ground their findings in qualitative understanding poses what Lofland calls a major contradiction between their public insistence on the adequacy of statistical portrayals of other humans and their personal everyday dealings with and judgments about other human beings:

> In everyday life, statistical sociologists, like everyone else, assume that they do not know or understand very well people they do not see or associate with very much. They assume that knowing and understanding other people require that one see them reasonably often and in a variety of situations relative to a variety of issues. Moreover, statistical sociologists, like other people, assume that in order to know or understand others one is well advised to give some conscious attention to that effort in face-to-face contacts. They assume, too, that the internal world of sociology--or any other social world--is not understandable unless one has been part of it in a face-to-face fashion for quite a period of time. How utterly paradoxical, then, for these same persons to turn around and make, by implication, precisely the opposite claim about people they have never encountered face-to-face--those people appearing as numbers in their tables and as correlations in their matrices! (Lofland, 1971:3.)

This returns us to the recurrent theme of matching the evaluation methodology to the problem. The highly informal, personalized environment of open education obviously lends itself to a more personalized evaluation methodology built upon close observer-student and observer-teacher interaction. Such a personalized evaluation is important not only for the insights it can generate but because *a personalized evaluation that takes the observer close to the data is the only evaluation research likely to be perceived as legitimate by program participants themselves.* To the extent that judging the quality of evaluation research includes judging its legitimacy and usefulness to program participants--and we would argue that this criteria should be central--then the matching of evaluation methodology to the nature of the program being evaluated is also central.

Finally, in thinking about the issue of closeness to the data, it is useful to remember that *many major contributions to our understanding of the world have come from scientists' personal experiences.* One finds many instances where closeness to the data made possible key insights--Piagets' closeness to his own children, Freud's proximity to and empathy with his patients, Darwin's closeness to nature, even Newton's intimate encounter with an apple. The distance prescribed by the dominant paradigm makes such insights derived from personal experience an endangered species.

8

Holistic vs. Component Analysis

Nowhere is the need to match methodology and problem more evident than in the dichotomy represented by holistic versus component analysis. Component analysis achieves its highest expression in classical Fisherian experiments using factorial designs, the most highly lauded of all evaluation designs (cf. Rossi, 1972:46). Experimental designs by their nature usually focus on some narrowly defined set of variables, at least one of which is the treatment. Causes are separated from effects, and both cause variables have to be carefully delimited and operationally defined.

Treatments in educational research are usually some type of new hardware, a specific curriculum innovation, variations in class size, or some specific type of teaching style. One of the major problems in experimental educational research is clear specification of what the treatment actually is, which infers controlling all other possible causal variables and the corresponding problem of multiple treatment interference and interaction effects. It is the constraints posed by controlling the specific treatment under study that necessitates simplifying and breaking down the totality of reality into small component parts. A great deal of the scientific enterprise revolves around this process of simplifying the complexity of reality. While this process is inevitable, it is also distorting. And it is the narrowness of focus in most experiments, with all their artificial controls and isolated treatments, that leads to the preponderance of "so what?" results, even on those rare occasions when significant differences in treatments are uncovered. The additional questions of the relevance of laboratory experiments for field settings only increases the distance between what is evaluated in most experiments and what actually happens in most classrooms or social action programs. Despite the dismal, disappointing, largely meaningless, and irrelevant (from the point of view of practitioners) results of thousands of educational experiments and quasi-experiments, the spokesmen for the dominant paradigm still argue that such designs are "the only available route to cumulative progress" (Campbell and Stanley, 1966:3).

Clearly there are questions of major import that do not lend themselves to experimental design or even less rigorous quantitative methodologies that focus on a limited number of narrowly defined variables. The simplified

world of variables, causes, and effects, in which the scientists of the dominant paradigm operate is alien to most teachers and change agents. Evaluations that are relevant and meaningful to the total context in which innovations occur need to include a holistic methodological approach built on the functioning, day-to-day world of program participants.

A holistic evaluation methodology is particularly crucial for holistic program innovations--like open education. Open education is an all-encompassing innovation. It involves not only changes in curriculum, materials, and methods, but also changed social relationships that affect the entire structure of the child's learning environment. Open education means new roles for teachers and learners, changed status arrangements in the classroom, a new set of norms, new expectations, and different criteria for evaluation. Interactions among students and the relationships between students and teachers are changed. Under conditions of such all-encompassing innovation it is impossible to specify what *the* treatment is. Moreover, it is impossible to carefully isolate and control component parts of open classrooms because the parts are so interdependent and interacting (cf. Patton, 1973).

To evaluate the meaning of open education as a holistic phenomenon requires a methodology that gets close to the classroom experience of children, a methodology of participant observation, in-depth interviewing, and careful descriptive detail that is subjective in the sense I specified earlier--the sense of discovering the meaning of the classroom experience from the point of view of the children and teachers.

A holistic evaluation methodology attempts to transcend the artificial conflicts in modern schools described by John Dewey in *The Child and the Curriculum*: "We get the case of the child vs. the curriculum; of the individual nature vs. social culture. Below all other divisions in pedagogic opinion lies this opposition" (Dewey, 1956a:5). A major component of this artificial conflict, for Dewey, was the division and specialization of subject matter in the curriculum. Academic divisions, he argued, are alien to the nature of the child:

> Again, the child's life is an integral, a total one. He passes quickly and readily from one topic to another, as from one spot to another, but is not conscious of transition or break. There is no conscious isolation, hardly conscious distinction. The things that occupy him are held together by the unity of the personal and social interests which his life carries along...(His) universe is fluid and fluent; its contents dissolve and re-form with amazing rapidity. But after all, it is the child's own world. It has the unity and completeness of his own life (pp. 5-6).

In contrast to the wholeness of the child's perceptions and experiences, "he goes to school, and various

studies divide and fractionalize the world for him" (p. 6). Dewey argued that in contrast to the school's methods of specialization and division, "the only significant method is the method of the mind as it reaches out and assimilates....It is because of this (specialization) that 'study' has become a synonym for what is irksome, and a lesson identical with a task" (p. 9):

> Abandon the notion of subject-matter as something fixed and ready-made in itself, outside the child's experience; cease thinking of the child's experience as something hard and fast; see it as something fluent, embryonic, vital; and we realize that the child and the curriculum are simply two limits which define a single process (p. 11).

Despite the totality of our personal experiences as living, working human beings, we have focused in evaluation research on parts, not only instead of wholes, but to the virtual exclusion of wholes. "We knew that human behavior was rarely if ever directly influenced or explained by an isolated variable; we knew that it was impossible to assume that any set of such variables was additive (with or without weighting); we knew that the complex mathematics of the interaction among any set of variables, much less their interaction with external variables, was incomprehensible to us. In effect, although we knew they did not exist, we defined them into being" (Deutscher, 1970:33).

While the radical critique of component analysis made by Deutscher in the last paragraph will be considered unacceptably extreme by most scientists, I find that teachers and practitioners voice the same criticisms about the bulk of evaluation research. Narrow experimental results lack relevance for innovative teachers because they have to deal with the whole in their classrooms. The reaction of these teachers to scientific research is like the reaction of Copernicus to the astronomers of his day: "With them," he observed, "it is as though an artist were to gather the hands, feet, head, and other members for his images from diverse models, each part excellently drawn, but not related to a single body, and since they in no way match each other, the result would be monster rather than man" (cf. Kuhn, 1970: 83). What teacher has not complained of the educational evaluation monster?

It is no simple task to undertake holistic evaluation, to search for the *Gestalt* in innovative classrooms and program innovations. The challenge for the participant observer is "to seek the essence of the life of the observed, to sum up, to find a central unifying principle" (Bruyn, 1970:316).

Again the work of Shapiro in evaluating innovative Follow Through classrooms is instructive. She found that test results could not be interpreted without understanding the larger cultural and institutional context in

which the individual child is situated:

> The relevance and appropriateness of the classroom and the test situation as locations for studying the impact of schooling on children requires re-evaluation. Each can supply useful information, but in both situations the evidence is situation-bound. Neither yields pure measures, and it is necessary to consider the type of school situation the children are in and their developmental status, as well as the social and sociological factors that determine or have determined the children's expectations, perceptions, and styles of thinking and communication with other children and adults. What may be an appropriate situation for assessing some groups may lead to misevaluation of others. The standard test, given under optimal conditions, may offer moderately valid estimates of competence for middle-class children (though every psychologist is aware of at least a few cases of gross misevaluation). Its adequacy and appropriateness may depend on unspecified built-in lines of continuity between middle-class cultural expectations and the demands of the test situation, rather than on intrinsic characteristics of the test itself. For lower-class children of different backgrounds there may be no comparable set of connectives, or the test situation may call for a type of response which is not valued in the child's cultural milieu. It is an old chestnut that psychological dimensions cannot be defined in terms of their physical equivalence; *psychologists who are trying to study the impact of different kinds of experience on different kinds of children must be able to shift their expectations and tools depending on the contexts in which they are working.* (Shapiro, 1973:541.)

Neither the holistic approach nor component analysis represents an omnibus strategy appropriate to all situations and problems. But in reaction to the dominance of component analysis as *The* Scientific Method in evaluation research this paper has emphasized the potential for more holistic evaluation strategies for holistic program innovations.

9

Process vs. Outcome Evaluation

The dominant scientific paradigm in evaluation research is preoccupied with outcomes. As with component analysis, the highest expression of this preoccupation is found in experimental designs. There is a pre-test, a treatment, and a post-test. The scientific observer enters the picture at two points in time, pre-test and post-test, and compares the treatment group to the control group on post-test measures. As already noted, such designs assume a single, identifiable, isolated, and measurable treatment. What's more, such designs assume that once introduced, the treatment remains relatively constant and unchanging.

While there are some narrow educational treatments that fit this description, more encompassing program innovations in practice are anything but static treatments. Frequently, by the time innovations are put into practice, they are already different than they appear in program proposals. Once in operation, innovative programs are frequently changed as practitioners learn what works and what doesn't, as they experiment and grow and change their priorities.

All of this, of course, provokes nearly unlimited frustration and hostility from scientific evaluators who need specifiable, unchanging treatments to relate to specifiable, pre-determined outcomes. Because of a commitment to a single evaluation paradigm evaluators are frequently prepared to actually do everything in their power to stop program adaptation and improvement so as not to interfere with their research design (cf. Parlett and Hamilton, 1972:6). The deleterious effect this may have on the program itself by discouraging new developments and redefinitions in mid-stream is considered a small sacrifice to be made in pursuit of higher level scientific knowledge. The arrogance and insensitivity of evaluators at such times--which are considerably more frequent than one might suspect--are all the more inexcusable when one considers that such interventions probably have already contaminated the treatment by affecting staff morale and participant autonomy.

Were some science of planning and policy/program development so highly developed that initial proposals were perfect,one might be able to sympathize with the desire of evaluators to keep the initial program implementation intact. In the real world, however, people and unforeseen

circumstances shape programs and initial implementations must be modified in ways that are rarely trivial. Nor is the task of program administrators and participants to shape their programs to the needs of evaluators. Rather the task of evaluators is to shape their evaluation methodologies to fit programs.

Under field conditions where programs are subject to change and redirection, the alternative evaluation paradigm replaces the outcome emphasis of the dominant paradigm with a process orientation. Process evaluation is not tied to a single treatment and pre-determined goals or outcomes. Process evaluation focuses on the actual operations of a program over a period of time. The evaluator sets out to understand and document the day-to-day reality of the setting or settings under study. Like the anthropologist, the process evaluator makes no attempt to manipulate, control, or eliminate situational variables or program developments, but takes as given the complexity of a changing reality. The evaluator tries to unravel what actually happens; he or she never takes for granted the implementation of a proposed treatment or innovation. The data of the evaluation are not just outcomes, but changes in treatments, patterns of action, reaction, and interaction. Under some conditions the initial and on-going observations of the evaluator can even serve as a source of program improvement--an impossibility under most controlled, static experimental designs.

In short, process evaluation requires sensitivity to both qualitative and quantitative changes in programs throughout their development, not just at some endpoint in time; it is built on subjective inferences in the sense that the investigator attempts to develop empathy with program participants and understand the changing meaning of the program in the participants' own terms; it requires getting close to the data, becoming intimately acquainted with the details of the program; it includes a holistic orientation to evaluation research, looking at not only anticipated outcomes but unanticipated consequences, treatment changes, and the larger context of program implementation and development.

10

Uniqueness vs. Generalization

The thrust of the dominant paradigm in evaluation research is a concern with discovery of scientific laws and theories. The Scientific Method is applied to uncover patterns of behavior; the ideal is to so specify and identify factors of social causation that the research scientist can explain 100 percent of the variance in social phenomena. The scientist in this instance seldom considers what a dismal world it would be if we could indeed account for 100 percent of the variance in human behavior.

The dominant paradigm is directed at producing generalizations. The assumption that this is the goal of Science is so deeply ingrained that it is virtually true by definition. I have never seen this assumption questioned in the literature on Scientific Methodology. Science *is* the search for generalizations.

Yet as human beings we place immense value on our individuality. Philosophers suggest that the greatest contribution of Western culture and civilization is the value it places on the individual. The rhetoric of educational innovation and social action programming is replete with references to reaching and serving individual clients. It strikes me that this emphasis on the individual has important implications for humanistic evaluation research.

Evaluation research studies in the tradition of the dominant paradigm report virtually nothing but norms, standards, surveys, and prediction equations. "But this very interest perhaps unduly distracts attention from the degree to which education is idiosyncratic as well as nomothetic. Teachers rarely feel they are facing merely 3 to 300 incarnations of points on a distribution; they hope they are educating Johnny Johnson and Suzy Smith. But, by those espousing the narrow definition (of Science, i.e. the dominant paradigm), dealing with the individual is usually considered an affair of art (medicine curing this patient) or technology (engineering building this bridge); the whole conceptual apparatus of science, along with its counterparts in educational philosophy and educational research, is often seen as inapplicable" (Dunkel, 1972:80).

In technical terms educational researchers sometimes recognize individuality when they discuss "disordinal interactions," i.e. treatments interacting with per-

sonological variables in educational experiments. This simply means that there may be some innovations that work better for certain types of students rather than showing across-the-board effects. Both Cronbach (1966) and Kagan (1966) have expressed the belief that the discovery method works better for some students than for others; some students will perform better with inductive teaching, and some will respond better to didactic teaching. Stolurow (1965) also has suggested that learning strategies interact with personological or individual variables.

Though such suggestions are hardly news to teachers (they know that different kids learn in different ways, though they don't always know how to take those differences into account in their teaching), disordinal interactions have rarely been uncovered in experimental research. Bracht and Glass (1968:449) report that while there are convincing arguments as to why one should expect disordinal interactions, "the empirical evidence for disordinal interactions is far less convincing than the arguments...." In point of fact, the actual search for disordinal interaction is rare--most researchers don't bother with the difficult statistical analyses necessary or don't measure relevant variables--and "the *molarity* (as opposed to the *molecularity*) of both personological variables and the treatments incorporated into many experiments *may* tend to obscure disordinal interactions which *might* be observable when both the variables and the treatments are more narrowly defined" (Bracht and Glass, 1968:451). Bracht and Glass (1968:452) conclude that "searching for such interactions with treatments as necessarily complex as instructional curricula may be fruitless."

In effect, Bracht and Glass prefer to dismiss the question rather than call into question the methodology that fails to find and predict individual differences. But for teachers, particularly teachers in innovative programs of open, informal, and humanistic education, the question will not go away. Indeed, for these teachers the central issue in the educational process is how to identify and deal with individual differences in children. Any serious and prudent observer *knows* that such differences exist, but experimental designs consistently fail to uncover them. Is it any wonder that practitioners find so much of educational evaluation useless and irrelevant?

Where the emphasis is on individualization of teaching or meeting the needs of individual welfare recipients --the 'clients' in social action programs, an evaluation strategy is needed that can take the individual into account. An evaluation methodology that takes the individual into account must be sensitive to uniqueness in both people and programs as well as similarities among people and generalizations about treatments. This is not a call for psychological reductionism, but rather an expression of what C. Wright Mills (1961) called "the sociological imagination"--a focus on the intersection of biography and history; attention to the interaction of the individual

and social structure.

The alternative paradigm of evaluation research can take account of the individual through its commitment to get close to the data, to be factual, descriptive, and quotive, i.e. to represent participants *in their own terms*. Lofland (1971:4), in describing such a humanistic approach to scientific research, argues that:

> ...this does not mean that one becomes an apologist for them, but rather that one faithfully depicts what goes on in their lives and what life is like for them, in such a way that one's audience is at least partially able to project themselves into the point of view of the people depicted.
>
> They can 'take the role of the other' because the reporter has given them a living sense of day-to-day talk, day-to-day activities, day-to-day concerns and problems. The audience can know the petty vexations of their existence, the disappointments that befall them, the joys and triumphs they savor, the typical contingencies they face. There is a conveyance of their prides, their shames, their secrets, their fellowships, their boredoms, their happinesses, their despairs....It is the observer's task to find out what is fundamental or central to the people or world under observation.

One of the effects of the overriding concern with finding generalizations in the dominant paradigm has been emphasis on ever larger samples, inclusion of an ever increasing number of cases in research studies, and the concommitant ever greater distance from and quantification of the data. The case study has fallen into disrepute in social science. Yet for certain types of questions, case studies in evaluation research are still very much needed. When the evaluation is aimed at improvement of a specific program, or when the information collected is for participants and not just scientists, and the concern is for individuals not just broad generalizations, then a case study approach that identifies uniqueness and idiosyncracies can be invaluable. Case studies can and do accumulate. Anthropologists have built up an invaluable wealth of case study data that includes both idiosyncratic information and patterns of culture. There is every reason to believe that the young discipline of evaluation research would be well served by a similar approach. More important is the likelihood that an in-depth case study would better serve program administrators and participants than the large-scale comparative studies aimed at finding similarities across program treatments. Not the least benefit of using the alternative paradigm is that the results are readily understandable to program participants and that their alienation from science and scientists is likely to be diminished--a humanistic consideration that has received little more than lip-service in most evaluation research.

11

Evaluation for Whom and for What?

The unanswered question underlying all of our discussion is for whom and to what end evaluative research is undertaken. It is a platitude in the evaluation literature that evaluative research should serve both scientists and practitioners. In reality, the needs of these two groups are frequently quite different. The dominant paradigm serves to delineate accepted and acceptable scientific practice. In terms of career considerations, personal legitimacy, and professional commitments, social scientists and educational researchers are best able to meet their needs by adherence to the prescriptions and standards of the dominant paradigm. The nature of funding in most major evaluative research reinforces this emphasis by rewarding grandiose designs, elegant sampling, and sophisticated quantitative methodological procedures. Such evaluations--frequently national in scope--focus on outcomes assessment and summative evaluation. *Such evaluations are virtually useless to practitioners in individual programs.*

Quite a different strategy is required where evaluation is aimed at serving and informing teachers and program practitioners about progress and functioning, areas of competence and confusion, attitudes, feelings, and practices which may be related to maximizing what the school or program has to offer. Evaluations that are to be useful to specific practitioners must be focused at the local level. They must include description and analysis of local settings. They must take account of what happens in programs on a day-to-day basis. We particularly need to be able to describe context, treatment, and outcomes in ways that are understandable, meaningful, and relevant to practitioners. The major value of this kind of program evaluation at this local level is its contribution to program development, not its labeling of successes and failures. The possibility for meaningful and useful feedback can occur only if evaluation research is tied to specific programs. It is also only at the local level that the decision of when to measure program impact can be made. National schedules for impact assessment almost invariably ignore variations in nature and degree of real program implementation.

While it is at the local level of immediate program evaluation that the alternative paradigm is most useful, this does not mean that it serves practitioners at the

expense of generating scientific knowledge of interest to the larger community. At the present stage of development of an interdisciplinary approach to evaluation research, with so little known about what constitutes a treatment or outcome and how evaluators can best measure these artifacts of social intervention, the alternative paradigm holds forth the promise of an accumulation of rich documentation that can serve well the larger goals of the scientific community.

12

Conclusion

I have outlined two paradigms of evaluation research. To facilitate analysis and discussion I have looked at these paradigms through a set of dichotomies: qualitative vs. quantitative methodology, validity vs. reliability, subjectivity vs. objectivity, closeness to vs. distance from the data, holistic vs. component analysis, process vs. outcome evaluation, and research for practitioners vs. research for scientists. In reality these are not dichotomies but continua along which evaluations and scientists vary.

As ideal-types, however, these dichotomies allow a kind of dialectical approach to consideration of the problem of competing paradigms. Though I have suggested only vaguely some possibilities for synthesis, my purpose has not been to undermine the dominant paradigm, but rather to plea for legitimacy for the alternative paradigm. Most important, I have argued that the evaluation strategy must be matched to the nature and needs of the evaluation problem and program setting.

Neither paradigm can meet all evaluation needs. The two paradigms have different strengths and weaknesses. It is my position that the strengths of the dominant paradigm do not justify its overwhelming monopolization of evaluative research and that the weaknesses of the alternative paradigm do not justify its current subordination.

Yet, as in any paradigm debate, great passions are aroused by advocates on each side. Kuhn (1970:109-110) tells us that this is the nature of paradigm debates: "To the extent that two scientific schools disagree about what is a problem and what a solution, they will inevitably talk through each other when debating the relative merits of their respective paradigms. In the partially circular arguments that regularly result, each paradigm will be shown to satisfy more or less the criteria that it dictates for itself and to fall short of a few of those dictated by its opponent....Since no paradigm ever solves all problems it defines and since no two paradigms leave all the same problems unsolved, paradigm questions always involve the question: *Which problems is it more significant to have solved?*"

Bibliography

Anastasi, A., "Preface." *Assessment in a Pluralistic Society: Proceedings of the 1972 Invitational Conference on Testing Problems*. Princeton, New Jersey: Educational Testing Service, 1973.

Becker, H., "Whose Side Are We On?" Pp. 15-26, in William J. Filstead, ed., *Qualitative Methodology*. Chicago: Markham, 1970.

Bernstein, I. and Freeman, H.E., *Academic and Entrepreneurial Research: Consequences of Diversity in Federal Evaluation Studies*. New York: Russell Sage Foundation, 1974.

Blumer, H., *Symbolic Interactionism*. Englewood Cliffs, New Jersey: Prentice-Hall, 1969.

Bracht, G.H. and Glass, G.V., "The External Validity of Experiments." *American Educational Research Journal*, 1968, 5:437-474.

Bruyn, S., "The Methodology of Participant Observation." *Human Organization*, 1963, 21:224-235.

_____. *The Human Perspective in Sociology: The Methodology of Participant Observation*. Englewood Cliffs, New Jersey: Prentice-Hall, 1966.

Bussis, A., Chittenden, E.A., and Amarel, M., "Methodology in Educational Evaluation and Research." Unpublished mimeograph, Princeton, New Jersey: Educational Testing Service, 1973. (Forthcoming in *Childhood Education*.)

Campbell, D.T. and Stanley, J.C., *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally, 1963.

Caro, F.G. (ed.), *Readings in Evaluation Research*. New York: Russell Sage Foundation, 1971.

Cronbach, L.J., "The Logic of Experiments on Discovery." Pp. 77-92, in Lee S. Shulman and Evan R. Keislar (eds.), *Learning By Discovery*. Chicago: Rand

McNally, 1966.

Deutscher, I., "Words and Deeds: Social Science and Social Policy." Pp. 27-51 in William J. Filstead (ed.), *Qualitative Methodology*. Chicago: Markham, 1970.

Dewey, J., *The Child and the Curriculum*. Chicago: University of Chicago Press, 1956.

_____. 1956b. *The School and Society*.

Dunkel, H.B., "Wanted: New Paradigms and a Normative Base for Research." Pp. 77-93 in Lawrence G. Thomas (ed.) *Philosophical Redirection of Educational Research:* The Seventy-First Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1972.

Filstead, W.J. (ed.), *Qualitative Methodology*. Chicago: Markham, 1970.

Gouldner, A.W., *The Coming Crisis of Western Sociology*. New York: Avon Books, 1970.

Horowitz, I.L. (ed.), *The New Sociology*. New York: Oxford University Press, 1964.

Hymes, D., "On Linguistic Theory, Communicative Competence, and the Education of Disadvantaged Children." Pp. 49-63 in Murray L. Wax, Stanley Diamond, and Fred O. Gearing (eds.), *Anthropological Perspectives on Education*. New York: Basic Books, 1971.

Kagan, J., "Learning, Attention and the Issue of Discovery." Pp. 151-161 in Lee S. Shulman and Evan R. Keislar (eds.), *Learning by Discovery: A Critical Appraisal*. Chicago: Rand McNally, 1966.

Lofland, J., *Analyzing Social Settings*. Belmont, California: Wadsworth, 1971.

Merton, R.K., *Social Theory and Social Structure*. Glencoe, Illinois: The Free Press, 1957.

Mills, C.W., *The Sociological Imagination*. New York: Grove Press, 1961.

Parlett, M. and Hamilton, D., "Evaluation as Illumination: A New Approach to the Study of Innovatory Programs," *Occasional Paper* No. 9. Center for Research in the Educational Sciences, University of Edinburgh, 1972.

Patton, M.Q., *Structure and Diffusion of Open Education: A Theoretical Perspective and An Empirical Assessment*. Unpublished Ph.D. dissertation, University

of Wisconsin, Madison, Wisconsin, 1973.

Petrie, H.G., "Theories Are Tested By Observing the Facts: Or Are They?" Pp. 47-73 in Lawrence G. Thomas (ed.), *Philosophical Redirection of Educational Research:* The Seventy-First Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1972.

Rossi, P., "Testing for Success and Failure in Social Action." Pp. 11-57 in Peter Rossi and Walter Williams (eds.), *Evaluating Social Programs: Theory, Practice, and Politics.* New York: Seminar Press, 1972.

_____. and Williams W. (eds.), *Evaluating Social Programs: Theory, Practice, and Politics.* New York: Seminar Press, 1972.

Scriven, M., "Objectivity and Subjectivity in Educational Research." Pp. 94-142 in Lawrence G. Thomas (ed.), *Philosophical Redirection of Educational Research:* The Seventy-First Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1972.

Shapiro, E., "Educational Evaluation: Rethinking the Criteria of Competence." *School Review,* November 1973, 523-549.

Stephens, J., *The Process of Schooling.* New York: Holt, Rinehart, and Winston, 1967.

Stolurow, L.M., "Model the Master Teacher or Master the Teaching Model." Pp. 223-247 in John D. Krumboltz (ed.), *Learning and the Educational Process.* Chicago: Rand McNally, 1965.

Strike, K., "Explaining and Understanding: The Impact of Science on Our Concept of Man." Pp. 26-46 in Lawrence G. Thomas (ed.), *Philosophical Redirection of Educational Research:* The Seventy-First Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1972.

Suchman, E.A., *Evaluative Research.* New York: Russell Sage Foundation, 1967.

Weiss, C.H. (ed.), *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, 1972a.

_____. *Evaluation Research: Methods of Assessing Program Effectiveness.* Englewood Cliffs, New Jersey: Prentice-Hall, 1972b.

Wirth, L., "Preface." Pp. x-xxii in Karl Mannheim, *Ideology and Utopia*. New York: Harcourt, Brace, 1949.

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

A Handbook on Documentation
Brenda Engel

An Open Education Perspective on Evaluation
George E. Hein

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