R E P O R T R E S U M E S

ED 015 793 FS 000 363

PROBLEMS OF EDUCATIONAL EVALUATION IN PROJECT HEAD START--SAMPLING, DESIGN, AND CONTROL GROUPS.

BY- MCDAVID, JOHN W.

PUB DATE 10 FEB 68

EDRS PRICE MF-$0.25 HC-$0.68 15P.

DESCRIPTORS- *PROBLEMS, *EVALUATION TECHNIQUES, EDUCATIONAL RESEARCH, *RESEARCH METHODOLOGY, *RESEARCH DESIGN, CULTURALLY DISADVANTAGED, *ACTION RESEARCH, DATA COLLECTION, CONTROL GROUPS, PROGRAM PLANNING, EXPERIMENTAL PROGRAMS, MEASUREMENT INSTRUMENTS, EARLY EXPERIENCE, CHILD DEVELOPMENT, HEADSTART, AERA,

CONTRARY TO THE OPINION OF MANY PEOPLE, PROJECT HEADSTART (HS) IS NOT A STABLE AND UNIFORM PROGRAM WHICH DEALS WITH AN EASILY DEFINABLE POPULATION. THERE ARE, THEREFORE, SEVERAL PROBLEMS WHICH EXIST IN CONNECTION WITH EVALUATIVE RESEARCH CONCERNED WITH HS. IN ORDER TO PROVIDE GUIDANCE IN PROGRAM PLANNING, THIS RESEARCH SEEKS TO DESCRIBE POTENTIAL RECIPIENTS OF HS ATTENTION AND POTENTIALLY USEFUL PROGRAMS, TO ESTABLISH SPECIFIC RELATIONSHIPS BETWEEN PROGRAM ELEMENTS AND POPULATION CHARACTERISTICS, AND TO EVALUATE SPECIFIC HYPOTHESES IN TERMS OF USEFULNESS. DUE TO (1) THE COMPREHENSIVE MULTI-DIMENSIONAL NATURE OF HS, (2) THE SIMULTANEOUS PURSUIT OF BOTH IMMEDIATE AND ULTIMATE IMPACT, AND (3) THE PAUCITY OF INFORMATION ABOUT THE DISADVANTAGED POPULATION AND ABOUT PRESCHOOL EDUCATION PROGRAM ELEMENTS, THE GREATEST INITIAL PROBLEM CONCERNED WITH HS EVALUATIVE RESEARCH IS A CONCEPTUAL ONE, THE FORMULATION OF QUESTIONS WHICH ARE PROPERLY "RESEARCHABLE." THE SECOND PROBLEM IS THAT OF METHODOLOGY, HOW TO SAMPLE AND TO DEVELOP MEASUREMENT INSTRUMENTS. SAMPLING PROBLEMS ARE ENCOUNTERED BECAUSE OF THE NON-RANDOM VARIATIONS IN HS POPULATIONS AND THE INACCESSIBILITY OF SUITABLE CONTROL GROUPS. THE THIRD PROBLEM IS THAT OF LOGISTIC DIFFICULTIES. IT IS NECESSARY FOR EVALUATIVE PROCEDURES TO BE UNOBTRUSIVE. TYPICAL CIRCUMSTANCES OF THE DISADVANTAGED HOME, LOW LITERACY LEVELS, AND THE PROBLEM OF RAPPORT BETWEEN DISADVANTAGED ADULTS AND MIDDLE-CLASS SCIENTISTS CONTRIBUTE TO THE LOGISTIC DIFFICULTIES ENCOUNTERED IN GATHERING RESEARCH DATA. THE FOURTH PROBLEM IS THE INTERPRETATION OF DATA IN HS EVALUATIVE RESEARCH. THIS PAPER WAS PRESENTED IN A SYMPOSIUM AT THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION MEETINGS, CHICAGO, ILLINOIS, FEBRUARY 10, 1968. (JS)

PROBLEMS OF EDUCATIONAL EVALUATION IN PROJECT HEAD START

Sampling, Design, and Control Groups

John W. McDavid, Ph.D.
Director of Research & Evaluation
Project Head Start

Paper presented in a Symposium at meetings of American Educational Research Association, Chicago, Illinois, February 10, 1968

PROBLEMS OF EDUCATIONAL EVALUATION IN PROJECT HEAD START

Sampling, Design, and Control Groups

John W. McDavid

Perhaps the greatest and most overwhelming difficulty in evaluating Project Head Start is that many people regard it as a stable, uniform, and easily-describable program dealing with a stable, uniform, and easily-describable population. This is simply not true.

Head Start is a massive social experiment, a piece of action research using a wide array of different samples of subjects, and carrying out a wide array of experimental manipulations under a general comprehensive umbrella lable, Head Start. Head Start grew out of the many discussions of social scientists which focused attention upon the existence of "cultures of poverty" that are self-perpetuating to the extent that values, attitudes, abilities, and habits of members of these subscultures are passed along from one generation to the next. Recognizing the implications of these ideas, educators and experts in child development formulated innovative plans for extensive programs of intervention into the early development of children born into disadvantaged subcultures to attempt to break this vicious cycle.

Since it is a social experiment, Head Start has been accompanied from its origin by a continuing program of evaluative research. Like most experiments, Head Start was initiated on the basis of a set of general hypotheses based in prior knowledge and theory about human development, education, and relationships between early childhood erperience and eventual adult behavior. Furthermore, since Head Start is an experiment, its planners did not expect total and unqualified success in attaining the program's objectives immediately.

In any experiment, the first observations of experimental consequences do not afford an oversimplified choice between abandoning the experiment as a failure or perpetuating it rigidly as a success. Instead, discoveries serve to redirect efforts along alternative routes, to focus attention in new directions, and to generate new ideas for further experimentation. Thus, Head Start's Research and Evaluation Office has planned its task accordingly: it has not attempted to provide immediate definitive answers about Head Start's ultimate success as a social experiment, but has instead framed a stepwise progression toward learning what kinds of intervention into early development are feasible, practical, and profitable in changing the intellectual and social skills, attitudes, and behavior of children and their families to enable them to produce greater contributions to their society and to enjoy a better mode of living. In brief, the steps in this progression of evaluative research include (a) the full description of the kinds of children and families with whom Head Start works, (b) description of the varieties of intervention programs which may be utilized by Head Start, (c) the establishment of specific relationships between program elements and population characteristics in terms of their consequental outcomes, and finally (d) the direct evaluation of specific hypotheses about programs and people in terms of their practicality and payoff, in order to afford future guidance in program planning. It would be unreasonable to expect immediate definitive answers about program alternatives and their success, since these answers must necessarily be preceded by investigations which establish the major dimensions of variation in people, programs, and consequences which need to be evaluated.

My comments in this symposium are intended to focus essentially upon problems of sampling and design, but in order to put them into prespective, I think I should first sketch briefly some peculiar problems associated with the overall context of what Head Start is and how it works. I have attempted to organize my thoughts into four categories: (a) conceptual problems, (b) methodological problems, (c) logistic problems, and (d) interpretational problems. The first catagory, CONCEPTUAL problems, is concerned with difficulties in formulating clear ideas and asking proper questions for evaluative studies. Within this context, I want to speak generally to you about what Head Start is and how its very commitments and objectives necessitate a formidable task of educational evaluation. The second category, METHODOLOGICAL in nature, concerns problems of instrumentation and procedure, and I will leave the discussion of this category to the other participants in this symposium. The third category, LOGISTIC problems, is indeed great in Head Start, since Head Start is a comprehensive and multidimensional federally supported program whose administrative organization places the greatest responsibility at the most remote local level. Head Start, nationally, is merely the formulation of a set of policy guidelines or boundaries of legitimacy in the expenditure of federally granted funds for local programs designed to achieve a specified set of objectives. Take note here: the objectives are specified, but the means by which they are to be achieved may vary considerably from one locale to the next. Most certainly, the populations with which these objectives are to be attained vary considerably from one area to another. Finally, I will speak only generally of the fourth category, problems of INTERPRETATION of findings, since I know that my colleagues on the symposium will also be directly concerned with this category.

Conceptual Problems

In any kind of research endeavor, the greatest initial problem posted for an investigator is that of formulating the proper questions. No amount of data collection can be of great research value unless a proper "researchable" question has first been framed. This has generated particular confusion in the early period of Head Start's existence.

First: Head Start is a Comprehensive Multi-dimensional Program. From its inception, Head Start was conceived as a program designed to influence the entire spectrum of a child's development; its objectives include improvement of the child's medical, dental, and nutritional status; improvement of his intellectual and academic skills and readiness for public school; improvement of his attitudes and feelings about himself and his relationship to other people and to society; improvement of social skills in relating to both his peers and to adults. Furthermore, Head Start is designed to influence not only the child, but also his family and home environment, and to influence his community context, including schools, neighborhoods, and other social institutions. Unfortunately, a large segment of the public has made the erroneous assumption that Head Start is merely a program to enhance school readiness and academic performance. To the extent that this wrong assumption has led investigators to over-concentrate their attention upon evaluating school readiness and intellectual status, much of the early research related to Head Start has been inappropriately narrow. It has been necessary for Head Start's Research and Evaluation Office to remind the public (including interested research investigators) of the broad scope of Head Start's definition.

Second: Head Start is Designed to Produce Both Immediate and Ultimate Impact.

The Head Start program is designed to produce certain changes in children, their families, and their communities which may be immediately apparent. These changes are assumed, on the basis of prior knowledge and theory of human development, to be likely to mediate later change when the child reaches elementary school, adolescence, and adulthood. But it is not assumed that there is perfect one-to-one correspondence between immediate consequences and long-range ultimate consequences. Certain immediate consequences of the Head Start experience (for example, diagnosis and correction of disease or defective eyesight or hearing, or social services to alleviate a harmful home situation) may have almost immediate consequences on the child's well-being and performance. Other immediate consequences (for example, improved self-regard and estimate of ability or improved attitudes toward authority figures) may be of less significance during dependent stages of early childhood and greater significance during later stages of greater independence. Some immediate consequences may be cumulative with later experience so that visible and noteworthy changes are not apparent for some time. It is likely that certain kinds of immediate consequence of the Head Start experience are transitory and are not durable unless subsequent conditions contribute to their maintenance. But under any circumstance, it is important to differentiate immediate effects of the Head Start experience from ultimate effects. Thus, it is crucial not to overinterpret research findings in either direction with respect to Head Start's short range immediate impact. This caution is perhaps most critical with respect to interpretation of intelligence test scores: not only are measures of intelligence subject to a variety

of interpretations (discussed later in this paper), but evidence has clearly established that measures of intelligence prior to age 5 are not highly correlated with measures at later ages (e.g. Thorndike, 1940); and more importantly, measure of IQ are of only limited value in predicting actual achievement and performance in academic or vocational situations (e.g. Thorndike & Hagen, 1955). The major implication of this issue is that final judgments of program impact of sweeping decisions about program policy and directives are clearly not warranted on the basis of studies of immediate impact of the Head Start experience in terms of a limited set of dimensions. Only long-range follow-up studies can supply the evidence necessary for such judgments, and Head Start is still too young for such studies to have been completed. It is important that all kinds of immediate impact be evaluated (including physical, attitudinal, social, and intellectual changes), and that these immediate consequences be followed into later stages of the child's development.

Third: Little is Known about the Children and Families whom Head Start Serves.

Until recently, the accumulated body of knowledge in the social sciences was based almost exclusively on the middle socio-economic classes. Little study, other than anecdotal descriptions and superficial statistical accounts, has been made of the characteristics of socially and economically disadvantaged people. When Head Start was created in 1965, little guidance was available to describe the kinds of people whom Head Start was to serve: their health and nutritional status, their attitudes and values, their habits and abilities were only rather vaguely identified. The very fact that sociologists had concluded that bounded cultures existed among the poor generated even further skepticism about the dangers of generalizing from experience with economically secure middle-class children

and families to the insecure and disadvantaged poor. It was necessary to keep an open mind, scientifically speaking, to the possibility that certain dimensions of relatively little consequence within the middle class take on new significance in studying the poor. Nutrition and health have traditionally been neglected variables in studies of the behavior of middle-class people, but there is reason to believe these dimensions may be important ones in dealing with the poor who are not comfortably well-fed, housed, and cared for medically. Patterns of family structure which are the rule for middle-class people are not necessarily typical of the poor or of particular minority groups (for example, the matriarchy of the American Negro family in which the father is often psychologically if not physically absent). Bilingualism, either in terms of a standard language system or in terms of sub-lingual dialects, is often associated with economic and social isolation. Even the availability of resources for communication and intellectual stimulation are drastically different for the poor (e.g., Horowitz & Rosenfeld, PRC, 1966). Thus, it is crucial to recognize that generalizations from research on middle-class children and families to apply to Head Start populations must be made cautiously, and unusual attention must be given to the equivalence of Head Start children with other groups when direct comparisons are made.

Fourth: Little is Known About the Elements of Preschool Education. Although programs of education for children under six have existed for two centuries, little knowledge has been developed to permit detailed description of specific curricular program elements. Descriptions have been made in terms of program philosophies, ranging from the permissive enrichment implications of traditional practice (e.g., Sears & Dowley, 1963; Alpern, 1966), through the "discovery-by-inquiry" approach (Suchman, 1960); and

the "learning-by-doing" philosophies (Montessori, 1967 ; Inhelder & Piaget, 1964), to more structured didiatic philosophies (Ausubel, 1961; Gray & Klaus, 1965; Deutsch, 1965; Radin & Weikart, 1967; and Bereiter & Engelman, 1966). But only a few of these descriptions have permitted detailed description and measurement of component elements in carrying out these philosophies.

Methodological Problems

Once properly framed questions for research have been posed, an investigator must proceed to collect data and make observations which permit him to answer such questions. He must gain access to a population (and usually only a sample of this total population) whom he can observe under specified conditions, and he must make observations according to carefully specified procedures (in the form of tests, observational records or rating scales, or procedures for coding the content of interviews and written records). In order to make use of these data, the observations should permit quantification so that they can be treated as scores in statistical analysis. Thus, the two major areas of methodological difficulty are (a) sampling, and (b) measurement instruments. Today my attention is focused on the former, and my colleagues on the symposium will address the latter.

Sampling Problems. In a laboratory, manipulations are planned according to an overall research design, but in field research, the research design must often be compromised to fit field conditions. Head Start was planned as a grass-roots community action program with decentralized responsibility for planning and execution. The resulting variability of decisions at the local level complicates the task of describing what Head Start actually is - - a preliminary task to planning an appropriate design for evaluation

of the program's effects.

Rarely are Head Start children randomly selected from among the population of eligible poor. With limited resources, some Head Start programs seek intentionally to serve the poorest and most severly disadvantaged families in a neighborhood, leaving a non-participant group of less disadvantaged children and families (Chandler, 1966; Wolfe & Stein, 1966; Johnson & Palomares, 1965; Coleman, 1966). In contrast, other Head Start centers predominantly serve families who have come forth voluntarily, and who are thus presumably more enthusiastic about the program and more concerned about the welfare of their children (Holmes & Holmes, 1966; Chandler, 1966). Families from foreign-language minorities, unusually mobile groups such as migrants, or with limited literacy or dire economic status, appear in some cases to be so far out of the mainstream of communication that they do not even become aware of the availability of a Head Start program (Wolff & Stein, 1966; Johnson & Palomares, 1965). Consequently, Head Start groups in different cities or neighborhoods are not always truly comparable to one another.

Such non-random variations among Head Start Centers complicated the design of an overall nationwide evaluation program by blurring and confounding the effects of different dimensions of variation among programs. For this reason, less ambitious investigations of circumscribed local programs are of great value even though the number of children observed may be small, and the range of variation within the program may be narrow. But in any kind of investigation, it is necessary to provide some reference point for interpreting what one observes. For descriptive purposes it may be useful to compare Head Start participants to middle-class preschoolers, but interpretations of such differences should be made cautiously.

It is clear that Head Start children begin their Head Start experience at a vastly different point of origin than middle-class children (NORC, 1965; Horowitz & Rosenfeld,1966). It appears that when they complete the Head Start experience they are more similar to middle-class children than they were originally, but still significantly different (Horowitz & Rosenfeld, 1966; Chesteen, 1966; Hodes, 1966; Hess, 1966; Eisenberg, et al 1966). But to gauge Head Start's impact some baseline or reference point must be established.

Perhaps the simplest is established by comparing each child to himself in a "before-and-after" design. A set of observations may be made before children are exposed to a Head Start program and then again afterward. But there are several reasons why such studies permit only tentative and equivocal interpretation. First, one cannot confidently assume that the observed pre- and post- differences are due directly to the intervening experience: Such differences might be reasonably ascribed (in any kind of measurement) simply to the passage of time and the normal development of the child, or to purely incidental and unaccounted for events during the interim, or (in the case of standardized tests) to practice effects and familiarity with test materials, or (in the case of observer ratings or judgments) to shifts in the judge's frame of reference over time. Furthermore, it is difficult to determine the most appropriate time for collection of "pre-" or "post-" data. It is almost impossible to collect data from children and families prior to the initiation of the Head Start program, because many of the techniques used to gather information are dependent upon observing the child in a group setting or testing him under specified and controlled conditions. Some interval of time is desirable to permit the child to accustom himself to new surroundings and stablize himself; on the other hand, there is evidence to suggest that

certain real changes in the child's behavior may occur within as little as two weeks after he enters Head Start (Pierce-Jones, et al., 1966). In addition, the very act of enrollment represents the first interventive step taken by Head Start, and it is one which focuses unusual attention upon the child and his family. Psychologists refer to the "Hawthorne effect" as a type of distortion of performance which comes when an individual realizes he is the target of special attention, and the direction of this effect is usually to "improve" performance by distorting behavior in a socially desirable, optimized direction (Homans, 1950). It is often difficult to isolate the Hawthorne effect sufficiently to permit clear interpretation of its distortion of data.

Under any circumstances, it is desirable to compare performances measures (including changes over time of intervening experience) for experimental groups against parallel control groups. The non-random variation of local Head Start groups makes it difficult to set specifications for proper comparison groups which are applicable to all Head Start centers across the nation, and as a result the national evaluation programs executed since 1965 (PRC, 1966; PRC, 1967; ETS, 1967) have lacked control groups for baseline data. Instead, the establishment of meaningful baselines for comparison can more readily be made in circumscribed investigations of a limited number of dimensions of program or population characteristics within a homogeneous set of Head Start classes. In follow-up studies to assess the residual effects of Head Start experience after the child has left the program (e.g., in kindergarten or first grade), it is obviously inappropriate to compare Head Start children to the remainder of their public school classmates without carefully matching the comparison sample on such pertinent variables as socio-economic status,

family structure and size, parental characteristics, age, and sex. Careless comparisons of non-equivalent groups can generate misleading data and unwarranted conclusions. Unfortunately, several studies conducted in public school settings have limited value because the Head Start groups and their correspondent non-Head Start comparisons were not equated on crucially relevant attributes (e.g., Wolff & Stein, 1966a; Coleman, 1966; Chesteen, 1966; Pierce-Jones et al., 1966). In establishing comparison groups to evaluate Head Start children, one-to-one matching or pairs of children is ideally desirable, but it is extremely difficult to match pairs on a number of pertinent characteristics simultaneously.

Logistic Problems

Certain logistic and mechanical problems are imposed on research in a social action setting, and these are often accentuated in the task of evaluating Project Head Start. The program's primary objective is service to children, families, and communities, and procedures for program evaluation must often take a back seat to program implementation.

Data collection procedures must be selected to be as unobtrusive as possible. Apart from the limited availability of testing procedures for preschool children, time-consuming tests which remove the child from the classroom for long periods are undesirable. The use of secondary data sources, such as teacher ratings, poses a number of difficulties. Not only are teachers not uniformly trained to make reliable and comparable objective judgments about children, but requesting them to do so for a number of children under their supervision is a tedious and unwelcome task. Without high degrees of task-involvement on the part of teacher judges, the quality and accuracy of data thus collected goes down sharply. Moreover, it is difficult to design any kind of a blind

study in which teacher ratings are employed as a data source because of the inherent fact that Head Start teachers are required to become highly familiar and closely involved with their children and their families.

Collection of data from parents is relatively difficult because of the typical circumstances of the disadvantaged home: the incidence of single-parent homes is high, parents work long hours, and rapport between the disadvantaged adult and the middle-class scientist is difficult to establish. Literacy levels are relatively low, so that complicated written questionnaires are inappropriate. Following a number of false starts with such procedures, the general direction of recent Head Start research and evaluation has now moved toward increased use of direct observation procedures, in which ratings and judgments are made by skilled and trained observers from sampled segments of time in the actual classroom setting. These procedures are costly, but the value of data thus gained justifies the investment of time, effort, and expense. When interviews are necessary, effort is made to use interviewers who are widely experienced with the disadvantaged or who have similar ethnic and racial backgrounds. It becomes necessary, of course, to take account of the characteristics of the tester or interviewer in interpreting the data collected.

One difficulty which follows from the use of observational procedures to collect data is that they are ordinarily standardized within the conventional classroom setting. Thus, they cannot be used comparably for parallel comparison groups of non-Head Start children who are not enrolled in an organized program. Either different or modified procedures must be employed, or evaluation must be interpreted only on the basis of internal comparisons among different kinds of Head Start (or similar)

programs for preschool children.

The respect for privacy and individual rights of children and their families may prohibit intensive exploration of particular dimensions of attitude and family relationships which would otherwise be of significant research interest. Although parental consent is obtained for all data collection in Head Start programs, within the context of federally sponsored research in a federally sponsored social program, the investigator must be particularly attentive to questions of undue invasion of privacy.

The problems of research design and sampling described here are not unique to Head Start research and evaluation; they have hindered the proper evaluation of educational impact for decades, and have even been cited as the primary reason for the critical shortage of research evidence on this issue (Jones, 1954; Hunt, 1961).