DOCUMENT RESUME

ED 097 958 PS 007 424

AUTHOR Zimiles, Herbert
TITLE A Radical and Regressive Solution to the Problem of Evaluation.
PUB DATE Jun 73
NOTE 12p.; Paper presented at the Meeting of the Minnesota Round Table in Early Childhood Education (Wayzata, Minnesota, June 1973)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Affective Tests; *Classroom Environment; Classroom Observation Techniques; Classroom Research; Cognitive Processes; Conceptual Schemes; *Early Childhood Education; Educational Objectives; *Evaluation Methods; *Evaluation Needs; Models; *Preschool Programs; Social Maturity

ABSTRACT

This paper reviews two major advances in preschool evaluation strategy that developed as a result of trying to evaluate Head Start, and proposes another evaluation approach. The first advance in evaluation procedure was to conceive educational objectives in terms of processes rather than products; that is, there was a shift from achievement tests to tests of cognitive process based on Piagetian problem-solving tasks. The second evaluation advance was to recognize the importance of comprehensiveness by extending evaluation content to include affective and social as well as cognitive processes. The alternative plan proposed in this report entails systematic and comprehensive evaluation of the child's school environment, to be followed by a theoretical analysis of the potential impact of his school experience. This approach represents a shift in emphasis from the assessment of impact on children to the assessment of the antecedent condition, the classroom environment. To implement such an approach to the evaluation of early childhood education programs, there is a need to explicitly formulate propositions regarding how and why preschool programs should work. On the basis of such a framework, methods must be devised for moving into a classroom and reliably describing, in quantitative terms wherever possible, the salient dimensions of its environment and its interactions. (CS)

A Radical and Regressive Solution to the Problem of Evaluation*

Herbert Zimiles
Bank Street College of Education

The irony of the title of this presentation stems from my observation that the more our current efforts to evaluate educational programs strive for relevance, the more invalid they become. Having reluctantly come to this conclusion, I propose that we radically change our methodological framework for evaluation. Let us examine the case for this proposal.

When Project Head Start was instituted, thereby vastly expanding preschool education, it was accompanied by a mandate to evaluate its effectiveness. The implication was that the program would stand or fall by this evaluation.

The evaluation of Head Start seemed precisely the situation which required the kind of comprehensive evaluation we at Bank Street College had been advocating and had begun to put into practice.* One of the guiding principles of our work has been the conception of schools as psychological fields, as environments which significantly influence children's psychological development--cognitive, affective and social--rather than as mere training grounds for academic skills. Our book, The Psychological Impact of School Experience (Minuchin, Biber, Shapiro, and Zimiles, 1969) reports the results of an effort to implement and test this point of view by systematic and empirical evaluation. The research was an intensive study of nine-year-old children who were attending very different kinds of schools. We examined the way in which these different educational experiences had affected the children's self-awareness, interpersonal skills, problem-solving patterns, group behavior, and other aspects of psychological functioning which relate to human development.

---

*Adapted from a paper presented at the Minnesota Round Table in Early Childhood Education, Wayzata, Minnesota, June 8-9, 1973.

The evaluation of Head Start, however, took a quite different, and more traditional turn. The first evaluation studies were conducted by psychometricians whose main concern was for the experimental design of the study. Few of the existing instruments had been standardized for use with young children, and since a quantitative evaluation requires a standardized test, the Stanford-Binet was almost automatically selected as the instrument to be used to evaluate the effectiveness of Head Start. Much more attention was given to problems of sampling, the designation of proper control groups, and appropriate methods of statistical analysis of the data. Nevertheless, questions inevitably arose regarding the relevance of Stanford-Binet items for an evaluation of the impact of preschool education and the search was on for intellectual measures whose content was closer to the teaching and learning which actually went on in preschool and which more accurately reflected the cultural values of the population under study. As a result, the priorities of standardization and quantification in the evaluation instruments were lowered and the criterion of content relevance was raised to a more central position.

The concept of relevance gradually broadened, and became increasingly sophisticated. Other measures of intellectual aptitude or achievement were added. Then a more significant change occurred. Largely under the impetus of the Piagetian rebirth, many investigators began to emphasize that preschool should be fostering the ability to think and function effectively on problem-solving tasks. The argument emphasized that preschools, especially those attempting to provide compensatory education, should be less concerned with training children to achieve specific skills or to learn specific academic content and more concerned with fostering cognitive growth--now that Piaget and Bruner and others had helped clarify what we meant by cognitive growth. Accordingly, evaluators were admonished to revise their assessment procedures still further and

focus on measures of cognitive process as well as cognitive achievement.

Each adjustment which defined criteria in greater breadth seemed to represent important progress; it meant that evaluators were beginning to see the fallibility of their simplistic criteria and that educators of young children were coming to grips with the fact that they were not merely concerned with training children to learn specific tasks. Program innovations such as the introduction of a "Piagetian curriculum" virtually dictated that evaluation criteria be defined in terms of cognitive process variables.

The next move forward, not surprisingly, was to extend the definition of educational objectives and evaluation criteria beyond the cognitive realm. The fact that many psychologists found this new domain an alien one is revealed by the reference to it as "non-cognitive." Thus, although the social and affective criteria were defined by exclusion they were, at least, beginning to be regarded as essential elements in a comprehensive evaluation battery.

Now, after less than a decade of intensive efforts to evaluate Head Start and the new programs in open education, two major advances have occurred: (1) educational objectives are being defined in terms of developmental processes rather than discrete products; and (2) the content of evaluation studies has been extended to include affective and social as well as cognitive processes.

While this amazing progress is to be applauded, one wonders how much advance in educational evaluation has actually been made. My own reservations are based on several considerations. Perhaps the most obvious concern is that when we examine the array of measures radiating from IQ and achievement tests to tests of cognitive processes and then to tests of social functioning and personality, we find a concomitant decline in validity. In attempting to measure cognitive processes rather than products, our use of problem-solving situations as opposed to conventional test items leads to a marked reduction in the amount of cognitive

behavior sampled, because it takes much more time to assess problem-solving behavior. While problem-solving tasks, on the surface, seem amenable to extensive analysis of qualitative features of performance, in reality, only a small number of behavioral characteristics can be categorized reliably. The net effect of introducing such new methods of assessment is to reduce the variability of scores which adversely affects both reliability and validity of measurement. Thus, problem-solving techniques have limited potential for yielding highly differentiating quantitative data, as compared with the wide range of scores and the high reliability of multiple-itemed intellectual aptitude tests which sample many domains. Personality measures are, of course, even less useful; at best, they have a degree of construct validity which cannot be understood in quantitative terms. It is hard to conceive of a single personality test which has the psychometric credentials to serve as a criterion measure in an educational evaluation.

Another disappointing note is that an increase in the breadth of assessment has not always been accompanied by a shift from product to process orientation. While conservation and other Piagetian cognitive attributes are replacing the learning of the alphabet in so-called innovative programs, such programs still seem just as concerned with training as those of the past. Conservation skills have merely replaced more traditional content in what remains a very traditional form of education. If children are to be drilled and trained, perhaps it would be better to train them in something that seemed useful to them, something which has face validity. Piaget uses the conservation paradigm, among others, to illustrate a mode and level of cognitive functioning. Whether or not a child conserves number may be quite revealing about his level of cognitive development, but it is not at all clear that a child who is trained to conserve is very different from one who has not been so trained.

If the recent reform in evaluation methodology has been distorted by many of those who have adopted a Piagetian approach, even greater errors of judgment have been committed in the name of personality assessment. I have received urgent phone calls asking for a good personality measure to be included in an evaluation battery in the same way that distributors are phoned by storekeepers regarding a new line of items they want included on their shelves. The fact that personality measurement remains one of the great unsolved problems of more than 50 years of research activity seems not to have penetrated those new converts who have suddenly recognized the value of comprehensive, developmental approaches to education. Their indiscriminate enthusiasm is not accompanied by an appreciation of the conceptual and methodological complexities involved in working with personality data. There is, therefore, every reason to be pessimistic about prospects for devising personality measures good enough to be used in large-scale evaluation studies. I have begun to believe that we have made an error in not taking Gordon Allport's (1937) call for idiographic measurement of personality more seriously. One of the problems with personality measurement is that different traits are differentially salient for different children (or adults). Across-the-board measurement of a particular trait generates a hodge-podge of data. The data gathered from those for whom the trait is salient may be quite telling, but the data obtained from the rest of the sample may have little or no functional significance.

During the days when we were struggling with the problem of evaluating Head Start, we were thwarted in our efforts to get Head Start teachers to tell us what their main objectives were and to describe how they proposed to reach them. The lack of readiness of educators to contribute to a substantial formulation of educational methods and goals has hampered evaluation studies. Finally, at the end of the school year, we turned to some articulate teachers in the Early

Childhood Center which Bank Street College was then operating in a poverty area and asked them to run down the list of children in their class, indicating for each child the areas of greatest growth during the preschool year. In almost every instance, these teachers singled out for consideration a facet of the child's personality or social behavior which had dominated his functioning in school and which had undergone change in response to their method of working with the child. But the attributes and context varied for each child. There was no question in the minds of the teachers who provided these data regarding the central role played by personality factors in the school lives of these young children, but it would have been impossible to capture the points they were making through the systematic application of a particular personality scale or inventory. Each child manifested a distinctive configuration of personality and social characteristics.

Another problem, well known to everyone but just as widely ignored, which bedevils those who seek a more relevant and comprehensive evaluation of school programs is the fact that a good deal of educational intervention is expected to have future rather than immediate impact. Yet evaluation research is so dominated by a mechanistic, push-pull outlook that we have learned to pretend that whatever findings show up immediately constitute the essential impact of an educational program. Such a perspective invites a narrow and superficial approach to education.

For all these reasons, none of them new, I cannot celebrate the long overdue move toward more relevant and more comprehensive evaluation. I have indicated that there are limits to the degree to which such goals can be attained and have observed that some of the notions of relevance and comprehensiveness have been misunderstood and distorted, thereby threatening to discredit the approach as a whole. I have also noted that comprehensive evaluation is severely limited unless

we are willing to assess the long-term impact of educational programs.

This very pessimistic analysis does not imply that the efforts described should be discontinued. We will not solve these important problems unless we continue to work at them. I can think of no more challenging research for a developmental psychologist than that of attempting to analyze the events of a preschool classroom in terms of their potential influence on the participating children, and then to devise an assessment of the children's characteristics which are hypothesized as being influenced. However, such work cannot and should not carry the label--or the burden--of evaluation because its findings, by definition, lack the infallibility and definitiveness we automatically associate with evaluation. When negative results are obtained they are much more likely to reflect the methodological weaknesses of the study than the failure of the educational program. The people working on such studies should not be constrained by the design requirements of evaluation, nor should they be required to carry the psychological and political burden of determining whether a program will stand or fall on the basis of a clearly inadequate study. Without the pressures of serving as an evaluator, researchers are likely to be less defensive, and more critical of their work and therefore freer to change and improve it.

If the evaluation of the impact of educational programs on children is to be discontinued because such evaluations are either too incomplete or, when they strive for comprehensiveness, invalid, then how shall programs be evaluated? The alternative plan here proposed simply entails systematic and comprehensive evaluation of the child's psychological school environment, to be followed by a theoretical analysis of the potential impact of his school experience. This would entail a shift in emphasis from the assessment of impact on children to the assessment of the antecedent condition, the classroom environment. Even those evaluation procedures which follow the current mode of focusing on the impact of

the program on the children are increasingly calling for a detailed description of the school environment. Their interest is primarily in more clearly defining the independent variable of an evaluation study. Many evaluation studies have reported outcome data on participating children without knowing with any degree of certainty or detail what the nature of the program was whose impact was being documented. Indeed, some evaluators make a virtue of such ignorance by claiming that they are unbiased by any prior exposure to the program whose impact they assess. During one of our evaluation studies of Project Head Start, we observed that many of the children whom we had extensively tested had hardly attended the Head Start program whose impact we were struggling to measure. It is equally absurd to assess the impact of a program without considering what actually went on in the program. Yet, most evaluators select their assessment instruments without firsthand knowledge of the program's way of operating. Apparently, evaluators view their task as a fishing expedition in strange waters; they cast the best nets available and hope for a good catch. The way in which the dependent variables which are being measured by the evaluation instruments are described makes it seem as though the measures have been chosen on the basis of a theoretical analysis of the actual educational phenomena to be evaluated, but in reality the measures are selected on the basis of convenience, availability, and a superficial judgment of relevance. As matters now stand, when one preschool program is reported as having "scored higher" in evaluation than another, my main conclusion is that the content of the arbitrarily chosen evaluation criteria more closely matched the transactions which took place in one program than the other.

Our inability to measure the impact of a program precisely or comprehensively is understandable in the light of existing methodological limitations, but these limitations do not apply to the task of conceptualizing and describing the program itself. Those who initiate and operate a program should be able to

describe what they are doing and what they are trying to accomplish. The task of describing and recording classroom interaction is of a very different order of magnitude from that of attempting to measure how a child's psychic organization and functioning has been affected by experiencing such an environment. It is a paradox that we have the responsibility and the capacity to describe and record the essential character of an educational program, yet do not do so; and at the same time, we do not know how to assess the impact of a complex set of experiences on the psychological functioning of a developing child, yet we persist in trying to do so.

But where are we in our evaluation if we simply document the nature of the program as it occurs but are unready to assess its impact on the participating children? We must carry our analysis of the program one step further. Just as it is the obligation of a program initiator and director to describe the nature of his program, so is it his responsibility to justify its usefulness on the basis of some specified conceptual framework. Any set of actions directed toward care and development of children is based upon an explicit or implicit set of propositions regarding the consequences of the proposed activities. Without a rational basis for its operation, a program does not deserve to be implemented.

Most educators operate on a largely intuitive level. Their conceptual framework is more implicit than explicit. The form of evaluation I am advocating requires that this framework become explicit. One of the greatest obstacles to progress in early childhood education is that formulation of the nature of the young child and his development is incomplete as is a conceptual scheme for educational programming in relation to our understanding of the child. If such an articulated theoretical framework existed, both in relation to the child and to an educational program for him, it should be possible to arrive at a set of

procedures for describing and recording educational environments and for analyzing such environments in terms of their potential impact on the participating children. Thus, we need a system that codifies observations of the adult models to which a child is exposed in school, the emotional climate of the classroom, the nature of the activities he experiences, the kinds of stimulation he receives, the values transmitted, and other related facets of the school environment that are likely to affect his development. In my view, this is the essence of educational evaluation and until we become better able to assess the impact of programs on children, our primary method of evaluating early childhood education programs should be to describe in great detail what they consist of and how they operate, and then hypothesize, on the basis of our theoretical framework, how a given program will affect children. While such a speculative approach to evaluation may lack the apparent advantages of current, preferred, empirical methods for validating a program, we are deluding ourselves, wasting time and effort, misinterpreting data and thereby subverting educational planning, by continuing to ignore the glaring deficiencies of empirical methods of evaluating educational impact and neglecting those activities of observation and theoretical analysis which are needed to shore up our conceptual framework for program planning. We need to observe children and programs much more than we do and we need to deal actively with the obligation to articulate and elaborate our conceptual framework. One of the reasons why assessment of impact has not progressed is the poverty of our thinking about children and programs. The more articulate we become about children and programs, the sharper and more effective will be our thinking about the assessment of impact. As already emphasized, I am not suggesting that efforts to assess impact should cease; on the contrary, they should expand, but not under the aegis of evaluation.

While the procedural changes I am recommending may seem radical, they are

not at all new, but simply describe how we now function most of the time. The proposition is that these procedures become codified. Most institutions and activities are evaluated in the fashion here recommended. We have very little systematic, experimentally controlled data regarding the efficacy of any of our most important activities or institutions. We do not know if going to a museum or library or concert really makes a difference nor do we have sound evidence regarding the value of taking a trip to Europe; yet we ungrudgingly spend large sums of money on such ventures. If we are selecting a camp for our child, we do not ask for data informing us about the average swimming speed improvement, nor would we be very much influenced by such data were it available. In our evaluation of the camp or the trip or the museum, we systematically examine the environment and analyze its potential for producing certain (usually multiple) desired consequences, and make our decision accordingly. As a matter of fact, I suspect that most of us, were we selecting a preschool for our children, would not place much stock in existing empirical validity data no matter how complete, but would instead base our evaluation on a visit to the school. Of course, it would be good if we could obtain sound, quantitative data regarding the value of all of the above-mentioned activities, but until such data are forthcoming, we would be wise to sharpen our methods for looking at and describing these institutions and developing our conceptual framework regarding how they function to produce particular outcomes.

To implement such an approach to the evaluation of early childhood education programs, we need to organize and elaborate our ideas and knowledge of young children, and formulate explicitly our propositions regarding how and why preschool programs should work. Given such a framework, we can move to the classroom for a reliable description, in quantitative terms wherever possible, of the salient dimensions which constitute its environment and its interactions.

We need to adopt this approach, not only because it will improve our methods of evaluation, but because of the impact it would have on current training and planning in early childhood education. It will foster an image of the classroom as a field, consisting of multiple interactions and dynamics which have a great variety of consequences. Evaluation of impact has had the effect of circumscribing the scope of a classroom. It fosters an approach to teaching in which the teacher works backward from the evaluation procedure; her concept of her goals and her methods become increasingly bound to the content of the evaluation instruments. If we need a jargon to describe these contrasting outlooks, we can term one mode of evaluation divergent and the other convergent. But most important, the procedure I am recommending places the focus of early childhood education where it belongs--on the study of children in school and the development of theoretical constructs for explaining the influence of their school experience.

References

Allport, G. Personality; a psychological interpretation. New York: H. Holt & Co., 1937.

Minuchin, P., Biber, B., Shapiro, E., & Zimiles, H. The psychological impact of school experience. New York: Basic Books, 1969.