DOCUMENT RESUME

ED 053 164 TM 000 681

AUTHOR	Dyer, Henry S.
TITLE	Issues in Testing.
PUB DATE	15 Jan 69
NOTE	13p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Achievement Tests, Cognitive Development, Cultural Differences, Culturally Disadvantaged, Disadvantaged Youth, Educational Programs, *Evaluation, *Intelligence Tests, Mental Tests, Motivation, Program Evaluation, *Socially Disadvantaged, *Testing, *Testing Problems

ABSTRACT

Certain concepts that are sometimes confused in discussions on testing socially disadvantaged children are clarified and a history of testing, beginning with Binet, is presented. Finally, five problems in using tests to evaluate educational programs for the disadvantaged are considered. (AG)

Issues in Testing

OECD Talk, Ford Foundation
January 15, 1969

Henry S. Dyer

Introduction

Before I get into the substance of this paper, I should like to make two brief preliminary remarks. The first has to do with the political aspects of educational evaluation which Mr. Halsey has asked us to consider. My paper does not deal directly with that matter, but since some experience has made me acutely aware of the problem, I think that, if you are attuned to it, you will hear plenty of political overtones in what I am about to say. My second observation has to do with a massive longitudinal study of disadvantaged children that ETS is launching under the auspices of the Office of Economic Opportunity. The study will focus primarily on 2000 children and will assess their cognitive, affective, and personal-social development from about age 3½ to about age 8. One of its purposes -- but only one -- is to provide an evaluation of the effects of Project Head Start. I shall make a number of references to the study in my paper.[1]

So much for the preliminaries. The paper I am about to read has three main parts. The first part attempts to clarify certain concepts that are sometimes confused in discussions of testing socially disadvantaged children. The second part takes a hop-skip-and-jump through the history of testing from Binet up to the present time. And the third considers five issues, or problems that tend to get in the way of attempts to use tests in the evaluation of educational programs for the disadvantaged.

I. Three Concepts

Two Sources of Confusion

Let me begin, then, by trying to disentangle the meanings of three terms that are often confused with one another. The three terms are (1) evaluation (2) testing and (3) the socially disadvantaged child. There is a tendency, in this country at least, to confuse the process of evaluation with the process of testing. Too many people have the notion that if you want to find out how an educational program is working, all you have to do is to give the children a standardized test as they enter the program, retest them with a parallel form of the test at the end of the program, and compute the differences between the pretest scores and the post-test scores to determine whether the program is effective. I submit that such a procedure does not tell you how well the program is working. Testing and retesting pupils is a necessary component of the evaluation process, but it is far from being the whole process.

There is also a tendency to fall into the confusion of using test results to identify the disadvantaged. If a child scores well below his age-mates on a reading test, for example, it is assumed implicitly that the low score indicates how socially disadvantaged he is. If somehow you raise his score, you eliminate his disadvantagement. This of course is absurd. It treats the independent and dependent variables as though they were one and the same thing, and skips over the crucial question of why a child is not doing well.

Meaning of Evaluation

In my view, the effort to answer this question of why some children do well and others do poorly is the substance of what we mean, or ought to mean, by evaluation

of educational programs. This point of view is not universal. There is, for instance, the notion that if you merely describe all aspects of a program and the conditions in which it operates, you have evaluated it. I grant that it is extremely important to gather all the descriptive data one can possibly get hold of, but until the data have been analyzed in such a way as to suggest why some children succeed and others do not, no evaluation has taken place. Let me hasten to add that I recognize that the answer to the question of cause will always be probabilistic, but I think that probabilistic answers are better than no answers at all.

Implicit in this notion of the evaluation process is another element that needs to be made explicit. If evaluation is the effort to answer the question why some children do poorly and others do well, then it is imperative to ask the further question: What do we really mean by doing poorly and doing well? And this raises two extraordinarily perplexing problems that have not had anywhere near the attention they deserve: (1) Who shall decide what constitutes good as contrasted with poor performance and (2) Once the decision is made, how do you go about detecting the difference between the poor and the good?

Meaning of Testing

This brings us to the matter of testing, for testing is the means by which, presumably, we can detect the difference between doing poorly and doing well. It is also, by implication, the means by which we ultimately define operationally what we really mean by good performance and poor performance. There is too much of a tendency to accept high test scores as good in themselves without considering the quality of the behavior that lies back of the numbers. In my view, testing, taken in its broadest possible sense, consists of any procedures by which individuals are

ordered along some continuum in accordance with their responses to standard situations. The standard situations may take thousands of different forms from multiple-choice questions to informal games that elicit interpersonal behavior. Any kind of situation can serve as a test, provided it is standard, that is, repeatable from pupil to pupil and from one occasion to another, and provided also that the pupil's responses to the situation can be observed and compared in some systematic and consistent manner.

Meaning of Socially Disadvantaged

Finally, what do we mean when we say that a child is socially disadvantaged? In my view a socially disadvantaged child is one who is being reared in a poverty-stricken family that suffers from discrimination in respect to jobs, housing, and access to the goods and services of the community. What I want to emphasize is that social disadvantagement in the present context is to be thought of as an attribute of the objective conditions in which the child grows up; it is not to be thought of as an attribute of the child himself. A blind or deaf child is disadvantaged, but not necessarily socially disadvantaged. A socially disadvantaged child may do very well on tests, but this does not make him any the less socially disadvantaged. On the average, of course, we know full well that socially disadvantaged children tend to score well below their socially more fortunate age-mates. A major problem in evaluating educational programs for such children is that of trying to determine how much of this average test deficit is to be explained by inadequacies in the program, how much by inadequacies in the social environment, and how much (and I underscore this) by inadequacies in the testing procedures themselves. It is for this reason that in undertaking any sort of educational evaluation, it is of first importance to keep clear the distinctions among three sets of variables: the social environment variables, the program variables, and the test variables. This is not easy to do.

II. Instant History of Testing.

One of the reasons I think it is useful to look back, however briefly, on the history of testing children is that such a backward look may help us avoid falling into the trap of supposing that this kind of testing is all problems and no solutions. To be sure, there are enormous problems in testing socially disadvantaged children that are badly in need of expert attention and research. But in the 64 years since Binet and Simon produced their first intelligence scale, we have in fact learned a great deal that can be put to practical use if we are sufficiently cautious and know what we are about. During the last six decades it is possible to identify five major strands in the development of testing.

The Binet Tradition

The first strand consists of tests in the Binet tradition, and it is of interest to note that from the beginning it was concerned with young children. The Binet-Simon scales of 1905 and 1908 were developed for children ages 3 to 11. By stretching a point, one might even make the case that they were concerned primarily with testing young children who were also "socially disadvantaged," so that the schools of Paris would be better able to meet their special needs. In short, Binet in 1904 was confronted by a very practical problem which was remarkably similar to the problem we are talking about today in 1969.

Binet had given a great deal of thought to the measurement of mental functions. He, like others before him, had experimented with various approaches to the problem with indifferent success. His genius consisted primarily in two fundamental contributions to the measurement of intelligence in children: the invention and develop-

ment of a set of psychologically complex exercises that differentiated between children who were acdemically successful and those who were academically unsuccessful, and the incorporation of these tasks into a normative scale based on the ages of the children tested. As you all know, the Binet scale, conforming to this normative approach, was expressed in units of mental age. To his lasting credit, Binet did not invent that rich source of academic confusion and exhausting psychological controversy, the notorious IQ.

Since Binet's time, innumerable mental tests for young children have been developed in the Binet tradition. In 1912, for instance, Kuhlman prepared a revision of the Binet scale which extended intelligence testing down to the age of three months; in 1916, Terman and his collaborators produced the first Stanford-Binet Intelligence Test, which begins with exercises for two-year olds; and as recently as 1967, there appeared the Wechsler Preschool and Primary Scale of Intelligence designed for children between ages four and six and a half. I shall not attempt to catalog all the Binet-type tests and the many variants of them that have appeared in the years between. My point is simply that, taken together, they represent a vast amount of very solid empirical work in identifying a great variety of developmental tasks for testing the mental functioning of children at the preschool and elementary school levels.

A characteristic of most of the tests in the Binet tradition is that they yield a single score on the child. They are based on the theory that intelligence is a single undifferentiated global entity that is expressed in many different kinds of performance. In the early days, and still to a remarkable extent, it was believed that this general ability to perform the tasks was principally, if not wholly, a

function of biological inheritance, and so the IQ, which was first introduced as a matter of clerical convenience, came to be enshrined as a kind of divine index of permanent expectation. People tended to forget that the tasks on general intelligence tests are tasks that children must learn to perform and that opportunities for learning how to do them may vary greatly from one child to another, as a consequence of the circumstances of their upbringing and schooling. This oversight on the part of many investigators has had the effect of fouling up any number of studies aimed at the evaluation of educational programs and systems.

The Factorial Tradition

A second strand running approximately parallel to the Binet tradition in testing the mental functioning of children has been another tradition which, for want of a better word, I shall call the factorial tradition. This takes the view that intelligence is not a one-dimensional glob as expressed in a single mental age or IQ score, but is multidimensional. That is, it suggests that a child can be high on one kind of ability, or factor, such as verbal fluency, and low on other kinds, such as speed of perception, or the ability to form concepts. The factorial tradition had its beginnings in the work of Charles Spearman in 1904 and it is still going on, attended by ample controversy among factor analysts. One of the important landmarks in this development was the Tests of Primary Mental Abilities produced by L. L. Thurstone and his wife in 1946 which provided scores on five factors: verbal, perceptual, quantitative, motor, and spatial. The important contribution of the factorial tradition which is still groping for a useful structuring of human ability is that it has emphasized the very great complexity of mental functioning. Its weakness, at this point in time, is that it has paid too little attention to the sequential

processes involved in the mental development of young children, and it has led some test users to regard the factor scores as measures of inherent abilities fixed at birth.

The Piaget Tradition

A third strand in the testing of young children began in 1922 with Jean Piaget's lectures on The Language and Thought of the Child at the Geneva School of Science. Piaget began his work, not by testing children, but simply by watching them and observing the thought patterns that emerged as they interacted with their environment. This represented a radical departure from both the Binet tradition and the factorial tradition. As Piaget himself has put it in a later work, the Binet approach to the assessment of mental development measured only the total "yield" of the child's thinking as a function of his age; it provided no information about the actual thought mechanisms that are operative at any given stage of development.[2] It is these mechanisms that we need to be able to observe and understand as precisely as possible if we are to meet the child where he is and shape instruction in such a way as to further his intellectual development. Not until quite recently have Piaget's ideas been incorporated into testing procedures usable in the classroom. The first effort I know of to bring the Piagetan approach into the schools themselves was one that we at Educational Testing Service began about five years ago in cooperation with the city of New York. There is no time here to describe that project in detail except to say that its principal goal was one of giving first-grade teachers themselves the means of observing the thought processes of their pupils in a natural setting so as to be able to understand the stage at which each child stood in his intellectual

development.[3] Tests based on the Piagetan concepts will play a prominent part in the massive longitudinal study of disadvantaged children that I mentioned above. One of the hopes of the study is that it will bring the educational process more closely into line with developmental processes of children.

The Achievement Testing Tradition

The fourth strand in the development of testing in this country is by far the most prominent: it is that which consists of standardized achievement tests. These came on the educational scene in the 1920's, and they have now proliferated to the point where the annual consumption in this country is probably well over 200 million. They vary enormously in quality and kind, and they are often misunderstood and misused in the assessment of educational operations. The tendency is to choose achievement tests according to the labels they bear rather than according to their content validity and their specific relevance to the kinds of behavior one hopes to measure.

As I see it, the main development in achievement testing that has taken place over the last 25 years or so is that of increasing the emphasis on questions that test for ability in problem solving and other types of reasoning and decreasing emphasis on questions that require simple factual memory and routine skills. This change in emphasis has had two important consequences: it has made achievement tests less distinguishable from aptitude and intelligence tests, and it has got the tests out of phase with most educational practice which is still characterized by a prime emphasis on memorizing facts and acquiring routine skills.

Testing Noncognitive Functions

The fifth strand in testing has been that concerned with attempts to measure the various aspects of character, personality, and social attitudes. This, too, has had a considerable history. As far back as 1911 the Psychological Bulletin began publishing periodic reviews of tests in this area as well as others. And the amount of high-level thinking and research that have gone into the problem ever since have tended to overshadow everything else that we have tried to do in the field of testing. The dimensions of the problem are well-illustrated by a single acerbic remark Oscar Buros makes in the introduction to his Sixth Mental Measurements Yearbook. Commenting on the number and distribution of the tests listed there, he notes that

> Personality, the area in which our assessment procedures have the least validity, is represented by the largest number of tests -- 196 or 16.1 per cent.[4]

One might say that this disproportion is a measure of the difficulty of the problem: the very fact that we have not yet been able to produce personality tests that we trust helps to explain why we produce so many of them. For we hope that somewhere, somehow we shall finally hit upon one or two that are sufficiently valid that we may dare to use them in the evaluation of educational programs.

I should add that the bulk of the work in testing noncognitive functions has been focused primarily on older children and adults. Interest in testing young children in this domain has been a fairly recent development, and it has been

particularly stimulated by concern for the affective behavior of the socially disadvantaged young. I might also add that an important feature of the ETS longitudinal study of the disadvantaged is that it will tie together all five of the traditional strands in testing that I have described and, by analyzing the interactions among the various measures over time, bring new illumination to the measurement of a very large number of both cognitive and noncognitive processes.

III. Five Issues

This brings me to the last part of this paper, the issues in testing disadvantaged children. There are, of course, hundreds of them, and I hope that, in my brief historical sketch of the situation, a few of them may have become visible. I shall conclude by listing five that I consider of major importance.

1. Probably one of the most severe problems in assessing disadvantaged children has to do with test motivation. It is difficult to get young children to pay attention to the testing situation and to put out their best effort. It is even more difficult to get disadvantaged children to do so. The consequence is that test scores of the disadvantaged young may seriously under-estimate or distort what they are capable of doing. And as they get older the amount of the error due to lack of motivation may be even more severe. There is some evidence to indicate that such children tend increasingly to perceive tests as symbols of the white middle-class culture that has rejected them, so they react by rejecting the tests. What are we going to do about this?

2. A second problem has to do with the popular conception of the meaning of the IQ as a number that unerringly reflects the child's so-called "innate ability" --

something that labels him for life as superior, average, or inferior -- something that is wholly independent of the conditions in which he has been reared. It is now beyond dispute that if you improve the conditions of learning you can improve the IQ, but the bulk of educational opinion hasn't caught up with this conclusion. What are we going to do about this?

3. A third problem has to do with the effects of cultural difference on test performance. Attempts to get around this problem by producing culture-free or culture-fair tests have been unsuccessful, because after you have drained off all the culturally loaded questions in any such test, you have a test that tells you essentially nothing about how the child is likely to succeed in any culture, whether it be his own or some other. What are we going to do about this?

4. A fourth problem is of a logistic nature. The testing of young children, whether we are concerned with their cognitive functioning, their self-image, their interactions with their peers, or what-not, usually requires that psychometrists who are well-trained in these matters shall observe these youngsters individually, one-by-one, in a face-to-face situation. There just are not enough of these psychometrists to handle the testing in any large-scale periodic assessment of children's development. What are we going to do about this?

5. Fifth, and finally, there is the problem of the experts themselves -- the researchers, the systems theorists, the cost-benefit analysts -- who are so fascinated by the testing of psychological theories and mathematical models that they tend to lose sight of the practical problems of testing children with a view to finding ways and means of helping them through measurement to find their way out of the ghettos. What are we going to do about this?

I am convinced that there are positive answers to all of these questions. I think we shall find a good many of the answers in our longitudinal study of disadvantaged children. I trust that more will emerge from the discussion that is about to ensue.

Notes

1. Educational Testing Service. Disadvantaged Children and Their First School Experiences; Interim Report, OEO Contract Number 4206. Princeton, N. J.: Educational Testing Service, February 26, 1968

2. Jean Piaget. Psychology of Intelligence. Patterson, N. J.: Littlefield, Adams, 1963. p. 153.

3. Educational Testing Service. Let's Look at First Graders: A Guide to Fostering and Understanding Intellectual Development in Young Children. New York: Board of Education of the City of New York, 1965.

4. Oscar K. Buros. The Sixth Mental Measurements Yearbook. Highland Park, N. J.: The Gryphon Press, 1965. p. xxx