



6. Evaluation of Evaluations

CONSTANCE KAMII *
DAVID L. ELLIOTT

ONE of the weakest aspects of the recent surge of activity in early childhood education has been that which was described in Part 3D, the evaluation of programs. This paper deals with some of the practical and theoretical reasons that we are not yet in a position to know what kind of program is most effective for what kinds of children. Some of the problems center around the nature of the data we gather as evidence of what children have learned and how to interpret such data. Other problems arise from the fact that short-term assessments may or may not stand the test of a longer-term view.

Comparisons among programs are made difficult because they often do not share the same goals and/or accept the same kind of data as a basis for judging program effects. This problem is compounded by our tendency to search for the one program or method that is "best" for all children and all teachers in spite of increasing evidence that this is a fruitless search. Some of these points and others will be elaborated in this article.

We will first deal with what was referred to in a previous section as *summative* evaluation, that is, the determination of the extent to which a program has achieved, or

seems to be in process of achieving, its objectives. We will then turn to *comparative* evaluation.

Summative Evaluation

As long as we remain within the context of summative evaluation of individual programs, evaluations can be evaluated in a fairly straightforward manner. Scriven (1967) has suggested that curriculum evaluation must include the following three "matches": (a) the match between program objectives and the content of instruction; (b) the match between the content of instruction and the evaluation instruments used; and (c) the match between the program objectives and the evaluation instruments. To the extent that any one of these matches is absent, curriculum evaluation can be said to have shortcomings.

Perhaps the most serious problem in the evaluation of early childhood programs is the absence of the third match. Many

* Constance Kamii, Curriculum Director, Early Education Program, Ypsilanti Public Schools, Ypsilanti, Michigan (on leave in 1970-71 at the University of Geneva and the International Center of Genetic Epistemology); and David L. Elliott

programs have a poorly conceptualized mixture of general and specific goals such as "cognitive development" and "language development"; knowledge of the alphabet and ability to count, to name colors and shapes, to use correct grammatical forms, and to follow directions; "self-direction" and "positive self-concept"; and many others. Not all these goals are represented in the evaluation design, and those that are represented are open to a wide range of interpretations.

It is desirable to use instruments that have been standardized on a sizable population. What typically happens in view of this desirability is that instruments believed to measure development or learning in areas similar to those included in program objectives get used because of their availability. Many of these instruments, however, were originally developed for other purposes such as the diagnosis and understanding of mental retardation or psycholinguistic and perceptual-motor difficulties. Examples of these include the Stanford-Binet Intelligence Scale, the Peabody Picture Vocabulary Test (Dunn, 1965), the Illinois Test of Psycholinguistic Abilities (Kirk, McCarthy, and Kirk, 1968), and the Developmental Test of Visual Perception (Frostig, 1963). These tests assess some kinds of learning, but not necessarily those which a particular program intended to achieve.

A rise in scores on available instruments does not necessarily allow the prediction of performance in areas not directly measured by the test. The familiar use of IQ scores illustrates this problem. A typical objective of an early childhood education program is better academic achievement in later elementary school on the part of "disadvantaged" children. The operational definition of this objective can be stated in terms of scores on selected achievement tests at the end of, say, sixth grade. In the case of a preschool program, defining the objectives (and the evaluation criteria) in this way means waiting six years to find out whether or not the program "worked." However, even since before the days of "accountability," it has usually been necessary to carry out summative evaluation more quickly (typically at the end

of a summer session or a school year). Therefore, the instruments used at the end of nursery school or kindergarten have been those that are known to predict performance in the sixth grade. As a result, the goal of higher academic achievement in sixth grade comes to be equated with higher IQ scores at the end of nursery school or kindergarten.

This restatement of goals is based on a very questionable premise. The fact that a preschooler's cognitive ability and development, as reflected in his IQ score, correlate highly with his later achievement scores does not necessarily mean that raising his preschool IQ score will increase his cognitive ability and hence later achievement scores. This is a bit like expecting the gas tank to contain more gasoline when the needle on the gauge is moved toward "full."

There are at least five different reasons for a child's IQ score to be higher on a retest following a period of instruction. First, he may have acquired test-taking skills, such as how to follow directions and how to produce the response called for (for example, pointing, making marks, and answering verbally). A child can also score higher on an IQ test if he is made to feel more comfortable in the test situation.¹ In these two cases, the higher of the two scores can be said to be the more accurate estimate of whatever it is that IQ tests measure. The increase is in the score, and not in what the score is supposed to represent. A third reason for gain in IQ score might be that the program emphasized abilities that are similar to those called for on the test. In this case, the test can be said to assess transfer of training. A fourth possibility is that the program taught the test items directly, in which case the test can be considered a criterion-referenced test. As with the first two cases, the gain in IQ score is due less to an increase in general intelligence than to additional learning of the specific abilities tapped by the test.

The test items in these situations stand

¹ Zigler and Butterfield (1968) demonstrated that it is possible to raise IQ scores of disadvantaged children 8-12 points simply by encouraging them to respond and by giving easy questions between difficult ones.

for themselves rather than representing a broader scope of development or learning. Finally, a program may actually contribute to the development of general intelligence, and thus to whatever makes the IQ test a valid predictor of later academic achievement. To what extent this is the case with any existing program and/or for any kind of pupils is a question requiring considerable research.

It should be noted that three tests were developed in the 1960's for the purpose of evaluating more directly how much children learned in compensatory preschool programs such as Head Start. These are the Basic Concept Inventory (Engelmann, 1967), the Preschool Inventory (Caldwell, 1967), and the Tests of Basic Experiences (Moss, 1970). To the extent that these tests were designed with many of the common goals of compensatory preschools in mind, they meet Scriven's third requirement of a match between goals and evaluation instruments better than the standardized tests mentioned above. Questions can still be raised, however, about the scope of developmental dimensions covered by these tests. A good deal more research is needed to determine whether or not the tests possess more validity than either predictive or developmental validity. In other words, we need to know the extent to which the items predict later performance in school (and if so, in which areas), and the extent to which they merely represent what early childhood educators think preschoolers should know before going to school.

Some programs report follow-up data in the form of scores on more traditional achievement tests given in the first, second, or third grade. It is true that high achievement scores in the early grades predict high achievement scores in high school. However, Kohlberg and Mayer (1970) recently reviewed the available longitudinal research and found that achievement in high school *as such* does not predict what happens after high school in job success or level of occupational prestige attained. College attendance does indeed make a difference, but high school graduates who do not attend college

do not do any better in income or occupational level than do dropouts. It seems advisable for compensatory education programs to conduct longitudinal research to find out the extent to which "making it" in school leads to "making it" socioeconomically afterwards.

Another set of facts reported by Kohlberg and Mayer is that "while early elementary IQ predicts later achievement, early elementary achievement does not predict later IQ nor does it predict later achievement any better than early IQ." In other words, intelligent children learn fast the things that are taught in school, but learning what is taught in school does not seem to make children more intelligent. In short, achievement test scores, too, must be viewed with caution both in the early elementary years and at the high school level.

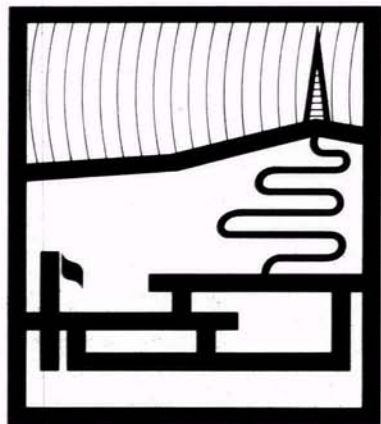
In summary, in reading evaluations of early childhood education programs we need to remind ourselves that summative evaluation has tended: (a) to use available instruments rather than developing more adequate ones; (b) to measure those dimensions of development that are easiest to measure and/or change; and (c) to have a short-term perspective rather than following through on the long-term consequences.

Comparative Evaluation

Attempts to compare the effects of different types of early childhood education programs entail all the above problems of summative evaluation, plus more. One set of additional problems is that certain dimensions of development (socio-emotional aspects in particular) are not included in the list of objectives of all the programs being compared. This means that certain dimensions will be omitted from the summative evaluation of certain programs.

Another problem is that the data bearing the same name are not necessarily comparable. For example, "curiosity," "language development," or "self-concept" is not always assessed under the same conditions. What is assessed with the child alone in a test situation is not necessarily comparable with what

SCHOOLS and cable television



Everything you need to know to plug your schools into cable TV in your community. Act now before it's too late!

How to obtain free public access channels

How to form public cable corporations

How to write a contract that ensures education's access to channels

How and what to negotiate

What costs are involved

What's predicted for the future of cable communications.

Order your copy now—only \$2.25 (stock #381-11968). Include payment or be billed on your institution's purchase order form. Discounts: 2-9 copies, 10%; 10 or more, 20%.

nea

National Education Association
Publications-Sales Section 135
1201 Sixteenth Street, N.W.
Washington, D.C. 20036

is observed in the classroom or at home. Likewise, for example, "language development" assessed by asking the child to "tell me all about this block" is not the same as what we obtain by asking him to explain how he wants to play a particular game.

However, the real problems involved in comparative evaluation are far more complex than the inclusion or exclusion of certain dimensions and the comparability of data bearing the same name. When comparative evaluation is carried out by an independent organization not involved in the operation of any program, it is often difficult to obtain agreement from individual sponsors as to what constitutes a fair evaluation.

As discussed in Part 2, there are many different views as to what is meant by human learning and development and the conditions under which they take place. These differences are reflected in the evaluation of programs, and there are consequently several different assumptions regarding what constitutes appropriate evidence of learning and

development. One of the factors complicating comparative evaluation is a failure to relate adequately the assumptions underlying evaluation (and hence strategies and instruments) to *any* of the major views of learning and development.

An example of the difficulty of comparative evaluation is the question of what span of time should elapse before the effects of a program can be assessed fairly. Some programs claim to be able to show effects in a few months. Others claim that their goals cannot be realized in less than, say, five years. For some theorists the ability to give as many correct answers as soon as possible constitutes evidence of an effective method of teaching. For others, speed of learning is less important than the building of a foundation for later learning.

Another factor that complicates comparative evaluation is the position various programs take with regard to school reforms. Some programs attempt to prepare the child for school as it exists today. Others take a

long-range view and proceed on the assumption that schools must and will change, and early education, too, must make the efforts necessary to hasten reforms. The programs based on these different assumptions cannot be compared in the same way with the same instruments.

All of this means that much research remains to be done on at least two related fronts before summative and comparative evaluation can be made adequately in early childhood education. One of these fronts deals with the conceptualization and evaluation of program objectives. If early childhood education is, in the long run, to contribute something of lasting value to the child's life, then short-term objectives must be evaluated in light of the long-term objectives of education. For example, the familiar question about whether socio-emotional objectives (and which ones) are less important, equally important, or more important than cognitive goals must be resolved in the long run by relating our social values to our theoretical assumptions concerning learning and development and the knowledge of what is possible to accomplish under what conditions. The evaluation of objectives will thus require an examination of empirical facts in relation to theories of learning, development, and education.

The second front on which research must be advanced is the development of better measurement techniques to determine what the various programs have accomplished in relation to their objectives. This means that the persistent old problems of measurement will have to be solved to improve the validity and comparability of the data we collect. One way to solve the problem is to develop tests which are administered under experimental conditions and standardized on a large population. However, since many dimensions of development do not lend themselves to customary testing conditions, it is necessary to develop standardized and comparable ways of observing children's behavior in their natural environment.

In our present efforts to obtain empirical facts, we may at best be employing

statistical theory as a substitute for psychological theories of learning and development. After using available instruments and statistical tests of significance, we are still faced with the problem of how to interpret statistical significance. The problem of measurement thus falls back on an examination of the theoretical assumptions underlying measurement. This means that more work is necessary to relate assumptions underlying measurement and evaluation to the major theories concerning learning and development.

In conclusion, we are far from being able to determine what kind of program is best for what kinds of children. Considerable further work is necessary both in the conceptualization of program objectives and in the development of instruments. Both these tasks will in turn rest on the formulation of theories concerning human development.

References

- B. M. Caldwell. *Preschool Inventory*. Princeton: Educational Testing Service, 1967.
- L. M. Dunn. *Peabody Picture Vocabulary Test*. Minneapolis: American Guidance Service, 1955.
- S. Engelmann. *The Basic Concept Inventory*. Chicago: Follett Publishing Company, 1967.
- M. Frostig. *Developmental Test of Visual Perception*. Palo Alto, California: Consulting Psychologists Press, 1963.
- S. A. Kirk, J. J. McCarthy, and W. D. Kirk. *The Illinois Test of Psycholinguistic Abilities*. Revised edition. Urbana: University of Illinois Press, 1968.
- L. Kohlberg and R. S. Mayer. "Preschool Research and Preschool Educational Objectives: A Critique and a Proposal." Cambridge, Massachusetts: Harvard University, 1970. Mimeo.
- M. H. Moss. *Tests of Basic Experiences*. Monterey, California: CTB McGraw-Hill Book Company, 1970.
- M. Scriven. "The Methodology of Evaluation." In: R. Tyler, R. Gagné, and M. Scriven. *Perspectives of Curriculum Evaluation*. AERA Monograph Series on Curriculum Evaluation No. 1. Chicago: Rand McNally & Company, 1967.
- E. Zigler and E. C. Butterfield. "Motivational Aspects of Changes in IQ Test Performance of Culturally Deprived Nursery School Children." *Child Development* 39: 1-14; 1968. □

Copyright © 1971 by the Association for Supervision and Curriculum Development. All rights reserved.