DOCUMENT RESUME

ED 030 948 EA 002 270

By-Stake, Robert E.

Toward a Technology for the Evaluation of Educational Programs.

Pub Date 67

Note-17p.; Pages 1-12 in PERSPECTIVES OF CURRICULUM EVALUATION, AERA Monograph Series on Curriculum Evaluation, edited by Ralph W. Tyler, And Others, Rand McNally & Co., Chicago, 1967.

Available from-Rand McNally & Company, Box 7600, Chicago, Illinois 60680 (Complete document 102 pages, $2.00).

EDRS Price MF-$0.25 HC-$0.95

Descriptors-Achievement Tests, *Curriculum Evaluation, Diagnostic Tests, *Evaluation Criteria, *Evaluation Techniques, Literature Reviews, Measurement, *Measurement Goals, *Program Evaluation

Identifiers-AERA, American Educational Research Association

Reflecting an increased awareness of the need for comprehensive curriculum evaluation, a monograph series has been initiated, focusing on major aspects of curriculum design and development. This introduction to the series defines curriculum evaluation as the collection, processing, and interpretation of two main kinds of data. (1) The objective descriptions of a curriculum's goals, environments, personnel, methods, content, and outcomes; and (2) personal judgments by the evaluator of the curriculum's goals, environments, etc. Available tests related to the evaluation of instruction seldom go beyond achievement testing. New techniques of observation and judgment of total curriculums are needed, with greater attention given to diagnostic testing, task analyses, and evaluation of goals. As reported in the growing literature on measurement and evaluation, special techniques employed in the behavioral sciences need to be utilized in curriculum evaluation. Through its sponsorship of the monograph series, the American Educational Research Association recognizes its obligation as well as its opportunity to cultivate a methodology for the evaluation of education programs. (JK)

AERA MONOGRAPH SERIES
N CURRICULUM EVALUATION

1

PERSPECTIVES OF CURRICULUM EVALUATION

Ralph Tyler
Robert Gagne
Michael Scriven

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
MONOGRAPH SERIES
ON
CURRICULUM EVALUATION

American Educational Research Association
1126 Sixteenth Street, N. W.
Washington, D. C. 20036

Perspectives of Curriculum Evaluation

Ralph W. Tyler

Center for Advanced Study in the Behavioral Sciences

Robert M. Gagné

University of California, Berkeley

Michael Scriven

Indiana University

$2.00

Rand McNally & Company
Chicago

Box 7600
60680

RAND McNALLY EDUCATION SERIES

B. OTHANEL SMITH, *Series Editor*

Printed in U.S.A. by Rand McNally & Company
Library of Congress Catalog Card Number 66-30794

Contents

Toward a Technology for the Evaluation of Educational Programs

"Let the buyer beware," declared the 19th-century individualist. What faith in human judgment! Then, as now, some "buyers" lacked the needed powers of judgment, but those who had them were expected to use them, and those who did not were said to deserve no concern. Yesteryear's spokesman for individualism advocated self-responsibility for one's choices and protested against government and consumer-collective action in the marketplace.

The marketplace in the last third of the 20th century will feature, along with provisions for sustenance, comfort, leisure, and longevity, a great array of products for the never-ending education of an inquisitive populace. Both government and private corporations have already initiated vast new production lines. Revolutionary curricula are emerging. Should the buyer beware? Can the buyer beware? What agencies are prepared to evaluate these educational products and programs? What steps should be taken to gain an understanding of an Operation Headstart or a school system designed by Litton Industries? We little understand the traditional operations of school systems. How can we understand the new?

How much we expect society to help the individual make decisions, in the marketplace and out, has changed greatly over the last 80 years. Most people now believe that government must not only protect against the grossly negligent and wanton, but must also license and standardize the conduct of legitimate business. Nongovernment agencies such as consumer organizations, professional associations, and producer self-regulatory bodies have been created to help provide information for judgment and decision.

How about the educational consumer? Can the teacher, superintendent, and curriculum coordinator choose wisely? Far too little information is now available. Little is known about the merit and shortcoming of products and programs. For excellence in

education we need excellent books and excellent teachers, but our methods of recognizing excellence are inadequate. For a few years, at least, there will be little quality control of goods produced by Research and Development Centers, by the growing curriculum-innovation projects, and by the newborn instruction industry. Much of the forthcoming educational output will be excellent, but not all. We grade the eggs a buyer cannot grade for himself and we legislate automobile safety standards. Yet far more crucially than eggs or automobiles, educational programs shape our future society. Should educational programs continue to escape formal evaluation?

ACCURACY VERSUS COMPLETENESS

"Let's call a spade a spade," declares a 20th century logical-positivist. What faith in perspicacity! To treat a spade properly we must recognize it as a spade. To specify the impact of an educational program we must be able to perceive impact.

Measurement specialists are proud of their perspicacity. "If it exists," they say, "it exists in quantity; and if it exists in quantity, it can be measured." It follows that if an educational program has an impact, that impact can be measured. Most specialists in educational testing and measurement believe they can do the job. The general public and most members of the educational profession presume that after having analyzed his data the "testing man" can state in precise terms the worth of a curriculum. The language of the Elementary and Secondary Education Act of 1965, Title I, implies that capability to evaluate is presently within our command. But the fluidity of our experiments and the bluntness of our tests deny us that capability. Neither quantity nor quality of impact is measured.

These are not, however, the greatest of our measurement problems: A spade is not just a spade. We do not have labels to identify each spade—and each educational program—so that it can be understood by label alone. Each needs ample description. Each differs from the others in a multitude of ways, and representation by title alone or by some composite score or rating leaves much of the story untold.

Our measurements are not perfectly accurate. We could devote ourselves to improving the precision of our instruments, but are there not higher-priority tasks? For the evaluation of curricula, I believe that we should postpone our concern for greater precision.

We should demonstrate first our awareness of a full array of teaching and learning phenomena. We should *extend* to this array our ability to observe and pass judgment. We should commit ourselves to a more *complete* description.

New techniques of observation and judgment need to be developed. In fact, we need a new technology of educational evaluation. We need new paradigms, new methods, and new findings to help the buyer beware, to help the teacher capitalize on new devices, to help the developer create new materials, and to help all of us to understand the changing educational enterprise.

PROFESSIONAL TOOLS AND TACTICS

It is not uncommon today for educational psychologists and measurement specialists to serve as advisors to evaluation projects. The inclination of these professionals, not surprisingly, is to use their most refined tools and techniques. Most of these tools and techniques were developed for differentiating among individual students, not for measuring the impact of an instructional program. Although differences in impact are indeed related to differences in student groups, curriculum evaluation and student evaluation require different measurement tactics.

Measurement consultants usually recommend specification of objectives in behavioral terms, experimental studies rather than status studies, and testing with instruments of empirically demonstrated reliability. Clearly these recommendations have their merit, but they can misguide evaluation efforts. J. Myron Atkin (1963) and Elliot Eisner (1966) have indicated how behavioral specification may disembody an educator's purpose. Lee Cronbach (1963) has indicated how a preoccupation with reliability can drain away an evolving test's content validity. Experimental controls are needed in the laboratory, regression equations are needed in the admissions office, and behavioral language is an essential consideration in test construction, but such techniques *may not* facilitate the conduct of an evaluation project.

Within the school, teachers and administrators evaluate their programs. Usually their purpose is self-improvement. When approaching the task in a formal way, they choose checklists and questionnaires as tools. Unfortunately, their inquiries are seldom validated, their attention to student achievement is negligible, and they seldom consider alternate ways of teaching. Still, these lay evaluations can be admired. They do attend to important facets of

the situation that are absent from the reports of measurement experts. New measurement tools and tactics should be devised for what is of obvious concern to the lay evaluators.

The official accreditation agencies and accreditation associations of the nation have not accepted the role of evaluator. They have established certain minimum standards. Each standard is believed to be related to quality education. When a school is rated, the extent to which standards are met is indicated—but the real worth of its educational program is not apparent in an accreditation report.

What strategies and tactics are needed for real evaluation? The writers of these first monographs, Ralph Tyler, Robert Gagné, and Michael Scriven, urge more attention to diagnostic testing, to task analyses, and to evaluation of goals. The approaches they offer are not completely new, but the attempt to bring them together for curriculum evaluation is all too new. Some of us see in these techniques the beginnings of a technology of evaluation. Our guess is that this technology will draw from instructional technology, psychometric-testing technology, social-survey technology, communication technology, and others; and that it will become a contributor to the understanding of evaluation in areas other than education.

The skeptical reader may respond that neither new tactics nor new tools are needed—that available tools used in the right way can do the job. Later, I will try to show why we should not expect certain common tools to be useful for evaluation, but first I want to specify what I mean by "curriculum evaluation."

A DEFINITION OF CURRICULUM EVALUATION

A curriculum is an educational program. It can be informally organized: what a craftsman teaches an apprentice; or formally organized: what is taught in an instructional film. A curriculum, defined in this way, could be a mere lesson, or it could be the curricular program of a comprehensive high school, or the entire educational program of a nation. A curriculum may be specified in terms of what the teacher will do, in terms of what the student will be exposed to, or—as Gagné does in this issue—in terms of student achievement.

Educational programs are characterized by their purposes, their content, their environments, their methods, and the changes they bring about. Usually there are messages to be conveyed, relation-

ships to be demonstrated, concepts to be symbolized, understandings and skills to be acquired. Evaluation is complex because each of the many characteristics requires separate attention.

The purpose of educational evaluation is expository: to acquaint the audience with the workings of certain educators and their learners. It differs from educational research in its orientation to a specific program rather than to variables common to many programs. A *full* evaluation results in a story, supported perhaps by statistics and profiles. It tells what happened. It reveals perceptions and judgments that different groups and individuals hold—obtained, I hope, by objective means. It tells of merit and shortcoming. As a bonus, it may offer generalizations ("The moral of the story is . . .") for the guidance of subsequent educational programs.

Curriculum evaluation requires collection, processing, and interpretation of data pertaining to an educational program. For a complete evaluation, two main kinds of data are collected: (1) objective *descriptions* of goals, environments, personnel, methods and content, and outcomes; and (2) personal *judgments* as to the quality and appropriateness of those goals, environments, etc. The curriculum evaluator has such diverse tasks as weighing the outcomes of a training institute against previously stated objectives, comparing the costs of two courses of study, collecting judgments of the social worth of a certain goal, and determining the skill or sophistication needed for students commencing a certain scholastic experience. These evaluative efforts should lead to better decision-making: to better development, better selection, and better use of curricula.

SOME LIMITATIONS OF AVAILABLE TESTS

Most contemporary evaluations of instruction begin and end with achievement testing. A large number of standardized tests are available. Many of these tests have been developed with appropriate attention to the *Standards for Educational and Psychological Tests and Manuals* (American Psychological Association, 1966) and to such well-considered guidelines as those in *Educational Measurements* (Lindquist, 1951, now in revision). It is important to our concern here to emphasize that these tests have been developed to provide reliable discrimination among *individual students*. Discriminability among students is important for instruction and guidance, but for development and selection of curricula, tests are

needed that discriminate among curricula. Different rules for test administration are possible, and different criteria of test development are appropriate, when the tests are to be used to discriminate among curricula.

For the usual standardized achievement test, the test author writes a large number of "general-coverage" or "general-skill" items. If certain content areas are unlikely to be encountered by many students, the author avoids them. Items on special content, even when valid, show up poorly on item analyses, and are weeded out. Since the items of a standardized achievement test are meant to be fair to students of all curricula, they are aimed at what is common to all. By intent, the standardized achievement test is unlikely to encompass the scope or penetrate to the depth of a particular curriculum being evaluated.

Items having a strong relationship with general intelligence usually look good in an item analysis. These items correlate highly among themselves and moderately with almost any achievement items. Since they *are* indirect measures of achievement which successfully predict subsequent performance, they are accepted by teachers and counselors as well as test developers. But indirect measurement of achievement is irrelevant, even offensive, to many curriculum developers and supervisors of instruction. They want to know *what* has been learned. They want to know what deficiencies remain in student understanding. The standardized test does not tell them.

Apart from clinical experience, our only current basis for interpreting most test performances is the frequency distribution of "total-test" scores collected from a norm group. Reputable test publishers have been reluctant to endorse subtest scores or to provide item response information. Clearly, individual-student decisions resting on responses to one or just a few items are questionable. Unlike the counselor, the curriculum supervisor does not concentrate on individual-student decisions. He must explain the variance among curricula. Test developers could help him by providing item data or, better still, by constructing separate subtests for each specific curricular objective. That would be a departure from current practice.

Please do not misunderstand me. I am not belittling our standardized achievement tests. I am favorably impressed with their usefulness for counseling students. But they are not equally useful for evaluation. I am dismayed by my colleagues who believe that these same tests can be used to satisfy the needs of the curriculum evaluator.

SOME NEEDED THINKING

The evaluator needs a battery of standard operating procedures. Procedures depend on criteria. Criteria depend on rationales. Rationales depend on theories. From evaluation theory to practice, new thinking is needed.

Regarding curriculum development, we need standard ways of translating aims and needs into practices. Our measurement and programmed-instruction specialists have developed taxonomies of objectives. Our classroom-learning-laboratory personnel have developed principles of instruction governing sequences of rules and examples, schedules for practice and review, hierarchies of understanding, etc. But there is no "compiler language," no grand scheme for deriving educational activities from given objectives. We need lesson-writing paradigms, including subroutines for helping an author maintain a pace, control reading difficulty, organize review exercises, discover inconsistencies, optimize redundancies, etc. Things like these, done today intuitively by authors and editors, should be done more explicitly with routine check on the quality of the materials written.

Whether accomplished by author and editor or by author and computer, the derivation of lessons should be examined on logical grounds. Today the evaluator lacks a rational procedure for checking the logic of the development of a curriculum. He needs ways to measure the correspondence between the intent of a lesson plan and the original goal. Does it require a thorough understanding of the subject matter? Should he employ a logician? We do not know.

As a part of this evaluation, communication integrity should be considered. Much of education includes the conveying of a certain message to a student audience. From the time a message is conceived until the students are exposed to it, a considerable transformation of the message is likely to take place. Does the author say what he wants to say? Does the teacher say what the author intended him to say? This concern applies whether the author is a subject-matter expert, e.g., a nuclear physicist, or the final transmitter, e.g., the physics teacher himself. Some writings are more illuminating than others, some homework problems are more pertinent than others, some demonstrations are more applicable than others. Some teachers use the right words but obscure the message, others refine and extend the message. To understand one important quality of a curriculum we must appraise the fidelity of its communications.

We need techniques for representing the perspectives held by

different people. Although they use the same language, different people see things differently. Do parents and the school board, consultants and the regular staff see things differently? Although two groups respond differently to a question, they may see the same merit in certain instruction. We need better devices for scaling perceptions of objectives. We need better procedures for processing judgments. At the beginning these procedures will not have the precision of an aptitude test or the elegance of an interest inventory, but even crude attempts to scale perceptions should be useful.

What are appropriate and inappropriate roles for the classroom teacher in curriculum evaluation? Can we capitalize upon the considerable ability of teachers to estimate which of two demonstrated teaching techniques is more likely to accomplish a particular long-term goal? Through training we could refine the teacher's powers of observation and estimation to make his contribution both technically sound and educationally valid. It is not unreasonable to conjecture that some day the primary role of the classroom teacher may be as a curriculum trouble-shooter, a conceptually oriented monitor, an evaluator—the essential link between the school's provision of a standard learning situation and the modification of it to accommodate the uniqueness of the student.

Several of the needs listed in this section call for psychometric thinking, the province of the psychologist. Other needs listed here and elsewhere call for help from the sociologist, the communications expert, the linguist, the philosopher, the anthropologist, and the economist. Can we find men of such pursuits to think with us as we develop our methodology of evaluation? I believe we must.

PRECURSORS OF A LITERATURE

Educational evaluation has not been without its champions. The social science literature includes many relevant works. A few will be cited here—more extensive coverage can be found in the bibliography.

Psychometric testing, for example, has been thoroughly discussed. For our purposes, the testing literature is nicely represented by *Educational Measurement* (Lindquist, 1951), with its farthest extension toward curriculum being Tyler's chapter on the measurement of learning. Among other fine writings on the evaluation of learning are those of Dressel and Mayhew, whose 1954 report is widely accepted as a textbook aimed at the evaluation of

course offerings, and portions of Ahmann and Glock's *Evaluating Pupil Growth* (1963). Many measurement projects deserve attention, but only two reports will be mentioned here: Project TALENT (Flanagan, 1964) and the National Assessment of Educational Progress (Tyler, 1966).

Techniques for the evaluation of teaching directly apply to curriculum evaluation. Prominent among writings in this area are Gage's *Handbook of Research on Teaching* (1963), and publications of McKeachie (1959) and Simpson and Seidman (1962).

Conant (1959), Gardner (1961), and Trump (1960) have made thorough but nontechnical evaluations of the nation's schools. Defining educational goals for the nation has been a continuing undertaking of the Educational Policies Commission (1959, 1961). More immediate instructional objectives have been the concern of Bloom (1956), Krathwohl (1964), Lindvall (1964), and their colleagues. The study of educational decision-making has been relatively neglected, but noteworthy are the works of Cronbach and Gleser (1964) and James (1963).

School environments, notably college environments, have been the focus of study by Astin (1961) and Pace (1965-66). Benson (1961), Carlson et al. (1965), and Mort (see Mort and Furno, 1960) have considered economic and social aspects of school systems. Questions concerning curriculum development have been discussed extensively by Taba (1962) and in a collection edited by Heath (1964b). On the general topic of innovation in education, Clark and Guba (1965), Miles (1964), and Pellegrin (1966) are frequently cited.

Innovation in measurement methodology is apparent in the literature. Methods well established in other branches of educational research have found applications in curriculum evaluation. Psychological scaling (Torgerson, 1958), Osgood's semantic differential (1957), and Flanders' interaction analysis (1961) are examples. The Damrin-Glaser tab-testing methods, adapted for group testing by McGuire (1966), seem to have particular promise.

As Britton (1964) found, much of the literature relevant to curriculum evaluation exists in impermanent form—office papers, conference handouts, etc. Many valuable illustrative pieces are virtually unknown because they were written only for the persons concerned with a particular curriculum. City-wide and state-wide evaluations get little attention outside their jurisdiction, but some have generated documents and instruments worthy of wider distribution. Some of the more noteworthy studies have occurred in Baltimore, and in the states of New York, Pennsylvania, Florida,

and California. Illustrative materials are sometimes available from consulting agencies such as the American Institute for Research; the Center for Instructional Research and Curriculum Evaluation at the University of Illinois; the Educational Testing Service; the Institute for Administrative Research at Teachers College, Columbia; and the Research and Development Center at UCLA.

THE CHALLENGE TO AERA

A professional organization sees no more clearly than its most sighted member, and seldom so well. Its actions usually serve more to consolidate rather than to extend, more to permit its members to tell of past deeds and future hopes than to propel them toward an institutional goal. So it has been with the American Educational Research Association.

It is possible for the more sighted members of almost any organization to become its officers. And so it has been with AERA.

In the early 1960's there were few independent sallies and little clamoring from the membership for the development of evaluation techniques. The officers of AERA, however, were then considering a possible impetus to evaluation efforts. They were aware that many new curricula were coming from such novel sources as National Science Foundation course-content improvement projects; that many special vocational programs were being initiated; that education had become a major instrument of war against poverty; and that the proliferation of programs now defies the local administrator's efforts to understand them on a personal basis.

Other professional organizations were also recognizing the need. The Association for Supervision and Curriculum Development, like AERA an affiliate of the National Education Association, devoted its Second National Conference on Curriculum Projects (Ammons and Gilchrist, 1965) to evaluation problems. The American Personnel and Guidance Association formed a subdivision called the Association for Measurement and Evaluation in Guidance. In 1965 a joint committee chaired by A. A. Lumsdaine and sponsored by AERA, the Department of Audio-Visual Instruction of the National Education Association, and the American Psychological Association prepared "Recommendations for Reporting the Effectiveness of Programmed Instruction Materials," a set of guidelines more general than the title suggests (Joint Committee

on Programed Instruction and Teaching Machines, 1963). Noteworthy publications have been provided by the American Association of School Administrators, the American Council on Education, the National Citizens' Council for Better Schools, the National School Boards Association, and the National Association of Secondary School Principals. The need for evaluation has not escaped the attention of any of these professional organizations, but none commands the broad research purview or the measurement skill to apply to that need. AERA does.

None of the publications cited in these two sections strongly encourages the belief that curriculum evaluation can be accomplished by currently available tests, checklists, and visitation routines. New tools and techniques are needed. With its involvement in the development and refinement of educational curricula, AERA is in a unique position. More than any other professional organization, it faces the obligation, and the opportunity, to cultivate a methodology for the evaluation of education programs.

THE COMMITTEE ON CURRICULUM EVALUATION

In 1964 President Lee J. Cronbach appointed an ad hoc committee to study possible AERA contributions. This committee, composed of N. L. Gage, Wells Hively II, John R. Mayor, and myself, reported early in 1965 that a number of activities were warranted. It recommended in particular that a regular committee be named, that conferences be sponsored by AERA, and that a series of monographs be published.

Acting upon this report and upon his own perception of educational affairs, President Benjamin S. Bloom in 1965 commissioned an AERA Committee on Curriculum Evaluation to develop guidelines for quality control—model evaluation procedures—to accompany the development and revision of educational curricula. Members of the 1965 committee were J. Stanley Ahmann, Leonard S. Cahen, Arthur Wells Foshay, Christine McGuire, Jack C. Merwin, Ernst Rothkopf, Richard A. Dershimer (ex officio), and myself. Harold Berlak and James P. Shaver were added in 1966 by President Julian C. Stanley. This committee, like its predecessor, concluded that guidelines limited to contemporary testing and inquiry procedures were inadequate; that special observation, data-reduction, and decision-making techniques were needed; and that AERA should encourage writing and discussion of theory and

rationales for such techniques. As a first project, this Monograph Series was proposed. It was approved by the AERA Board of Directors early in 1966.

AERA, of course, has no writers of its own. The Monograph Series was created to attract contributions from members and nonmembers alike. Many disciplines should be represented. A distinguished educator, a distinguished psychologist, and a distinguished philosopher have contributed to this first issue. It is expected that economists, social anthropologists, communications specialists, school administrators, and classroom teachers will be among the authors of future issues.

Some issues will contain several monographs; most perhaps will be devoted to a single monograph. Attention will range across such diverse topics as decision-making, educational goals, innovation, merit in teaching, merit in textbooks, the measurement of change, content validity, the politics of education—in short, to any topic that contributes to the scholarly study and technical practice of evaluation in education.

This Monograph Series is not a new professional journal. It will be published aperiodically, to meet a current need. It will be continued only as long as the priority of the need remains high. The Series will exist as a medium for writings too lengthy and too elaborate for journal publication. It will include discourse. Some of this discourse will be speculative, some may even be supplicatory. Although some of the contributions will be theoretical and abstract, the ultimate purpose of the Series is to serve the practitioner. The primary criterion for acceptance of a manuscript will be whether, in the long run, what the author has to say will facilitate the development of palatable, comprehensive, and dependable evaluation procedures.

At this point, we do not know what directions this Monograph Series will take or what services it will render. Will it aid the curriculum developer? Will it ultimately help the buyer beware? We are convinced that the purposes of evaluation should be reconsidered, that our resources should be inventoried, that new models of evaluation should be proposed, and that new tactics should be discussed. The monographs in this Series should serve these ends.

ROBERT E. STAKE