DOCUMENT RESUME

ED 109 169 TM 004 616

AUTHOR	Behrman, Edward H.; Evans, William
TITLE	Evaluating the Effectiveness of Administrative Training Products.
PUB DATE	[Apr 75]
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D. C., March 30-April 3, 1975)
EDRS PRICE	MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS	*Administrator Education; Comparative Analysis; *Curriculum Evaluation; Evaluation Methods; Formative Evaluation; Models; *Program Effectiveness; *Program Evaluation; Summative Evaluation

ABSTRACT

This paper clarifies the distinction between evaluating the effectiveness of student curriculum products and evaluating the effectiveness of administrative training products. Four product evaluation designs are described and critiqued along eight interrelated dimensions: the definition of product effectiveness; criterion measures of effectiveness; the determination of users' pre-treatment levels of ability on the criterion; types of comparisons; sample size and comprehensiveness; the determination of acceptable criterion performance; confounding variables; and intervening variables. Finally, a practical yet legitimate strategy for measuring effectiveness of administrative training products is proposed. (Author)

EVALUATING THE EFFECTIVENESS OF ADMINISTRATIVE TRAINING PRODUCTS

Edward H. Behrman
Research for Better Schools

William Evans
University of Pennsylvania

Presented at the annual meeting of the American Educational Research Association, Washington, D. C., 1975.

Originally entitled "Evaluating the Evaluation of Product Effectiveness."

EVALUATING THE EFFECTIVENESS OF ADMINISTRATIVE TRAINING PRODUCTS

Edward H. Behrman
Research for Better Schools

William Evans
University of Pennsylvania

Educational R&D, spurred by the creation of the labs and centers in the past decade, has created a proliferation of new products in the educational marketplace. Product evaluation has emerged as an area of considerable importance in the measurement field, as both the developer and the funding agent seek to assess the worth of the emergent product. Although product evaluation strategies have become more sophisticated in the decade, they nonetheless have been generally built to assess curriculum products designed for student populations (e.g., Bloom, Hastings, & Madaus, 1971; Gagne, 1967; Grobman, 1968; Scriven, 1967; Tyler, 1967). Such strategies are suggestive but less than helpful when attempting to evaluate other kinds of educational R&D products, such as those designed to train school administrators in aspects of educational management. Examples of this latter type of educational product include varied prototypes now under development, such as those designed to train administrators in project management, curriculum selection, curriculum evaluation, cost-effectiveness analysis, needs assessment techniques, etc.

The purpose of the present paper is to make clear the distinction between evaluating the two types of products and to propose one possible method for evaluating the effectiveness of administrative training products. In doing so, the paper will also highlight the difficulties, conflicts, and trade-offs encountered in determining product effectiveness of this latter group. The paper appears most useful to evaluators and developers,[1] who should be able to make valid statements about product effectiveness before requesting continued funding for product dissemination. It may also be helpful to funding agents, who must make dissemination decisions.

We should stress at the outset that the argument identifying the evaluation of product effectiveness as a summative rather than formative function is persuasive, but only to a point. Scriven (1967) has already noted that there is no absolute cut-off between formative and summative phases. The formative evaluator is often called upon to produce evidence of product effectiveness in order to obtain funding for continued development or dissemination. Therefore, concern over product effectiveness may be justified during both phases of evaluation, although the shape and scope of the effectiveness assessment is determined by both the nature of the product and its developmental status.

The method of the present study involves comparison of the formative evaluation strategies employed in the assessment of four educational R&D products, one being drawn from the evaluation of a student curriculum product and three being drawn from the evaluations of administrative training products. These four strategies are described and then critiqued along 8 interrelated dimensions: (a) the definition of product effectiveness; (b) criterion measures of effectiveness; (c) the determination of users' pre-treatment levels of ability on the criterion; (d) types of comparisons; (e) sample size and comprehensiveness; (f) the determination of acceptable criterion performance; (g) confounding variables; and (h) intervening variables. Although determination of product effectiveness is only one aspect of overall formative

[1] Throughout the paper, "developers" refers to any personnel employed by the product development agency.

evaluation procedures, it is the sole focus of this paper. Of course, other activities must precede the collection of product effectiveness data -- such as "debugging" content and instructions.

The data presented were gathered as part of on-going evaluation efforts at one of the educational labs. To give the reader a better idea of the context in which this paper deals, let us briefly describe each of these product evaluations. The student curriculum product is an individualized science curriculum. The administrative training products include one designed to train school managers in proposal development; another designed to train school managers in curriculum evaluation; and the revised prototype of the curriculum evaluation product.

THE FOUR EVALUATION STRATEGIES

1. *Student curriculum: science.* Students received instruction in the regular school setting throughout one academic year. A pretest/posttest, comparison group design was used to determine how well students in the individualized curriculum achieved two prespecified goals of the product (science achievement and attitude toward science). Students from three pairs of matched schools (n = 636 for pretest and 615 for posttest) were administered the developer-constructed achievement and attitude measures in the fall and again in the spring. The same form of each measure was used in pre- and posttesting. Students were tested as individuals but scores were grouped together by grade level and by school for data analysis. Both gain scores within schools and comparison scores between schools were computed (see Evans, 1973).

2. *Administrative training: proposal development.* School administrators from six districts were trained on-site in the district by another member of their own staff over a three-week period. A pretest/posttest, single group design was employed to evaluate whether administrators achieved the prespecified goals of the product (ability to manage a proposal development project). Content-recall was measured by 10 multiple-choice items for each lesson, with the same set of items serving as both pre- and posttests. In addition, simulation tests of performance -- actually exercises contained within the instructional lessons -- were used as additional, posttest-only measures. These simulations required learners to apply what they had learned to a hypothetical situation of the developer's creation. Significant gain scores on the mastery tests and subjective judgments by developers of quality work on the simulations were indices of acceptable criterion performance. Each of the 35 administrators was tested as an individual and then all scores were combined into a single group for data analysis (see Evans, Note 1).

3. *Administrative training: curriculum evaluation.* School administrators from two districts were trained on-site in their own districts by the developer during two-day sessions. A posttest only, single group design was used to determine if administrators could demonstrate achievement of the prespecified product goal (ability to initiate, plan, and monitor a curriculum evaluation project). Simulation tests of performance comprising exercises and worksheets included in the instructional materials were reviewed by developers to yield subjective ratings of quality performance. These simulations required learners to apply what they had learned to a hypothetical situation of their own creation. Six administrators from each district com-

prised the sample ($\underline{n}$ = 12). Since administrators from the same district worked collaboratively to complete exercises and worksheets, only group performance was scored (effective $\underline{n}$ = 2) (see Behrman, Note 2).

4. *Administrative training: curriculum evaluation (revised prototype).* Using the revised prototype, administrators were again trained how to manage a curriculum evaluation project. This time, though, they trained themselves on-site in their districts without developer support. Because the exercises now required the administrators to apply what they had learned to an actual, on-going evaluation project, training continued intermittently over several months. Again, a posttest only, single group design was used to determine if administrators could attain the prespecified goal. Performance was reviewed as in the earlier prototype, except that administrators applied learning to real rather than hypothetical situations: thus the tests were work samples rather than simulations. Thirty-nine administrators from four districts (working on seven separate projects) comprised the sample. Since project groups worked collaboratively, only group performance was scored (effective $\underline{n}$ = 7) (see Behrman, Note 3).

The four evaluations described above represent a fairly wide range of strategies. The evaluation of the science product seems to be a rather typical assessment of student curriculum product effectiveness (with the possible exception of employing specially-developed instruments). It is well-suited for the one-way ANOVA design and contains samples large enough to permit powerful statistical inference. Generalizations to similar students in similar schools are possible.

Such is not the case with the three administrative training products, which do not fit so neatly into the traditional evaluation model. That they fail to is not necessarily an indictment against the quality of these latter evaluations; in fact, it is a goal of this paper to show why attainment of the traditional design is so difficult (and may not even be desirable) when evaluating administrative training products. Therefore, the next step in our discussion is to critique the evaluation strategies described above along each of eight interrelated dimensions

CRITIQUE ALONG EIGHT DIMENSIONS

Definition of product effectiveness. Invariably, effectiveness was defined as the match between prespecified product goals and observed learner performance; that is, all four evaluations were explicitly of a "goal-full" rather than "goal-free" nature. The science curriculum specified its goals in rather broad terms (science achievement and science attitude). While the administrative products also specified broad goals, these were analyzed into sub-goals or objectives for measurement purposes. For example, the curriculum evaluation product measured the following objectives for units (or tasks) 1 and 2:

1. ability to construct an evaluation purpose statement
2. ability to develop an overall evaluation design
3. ability to specify evaluation instruments and subjects

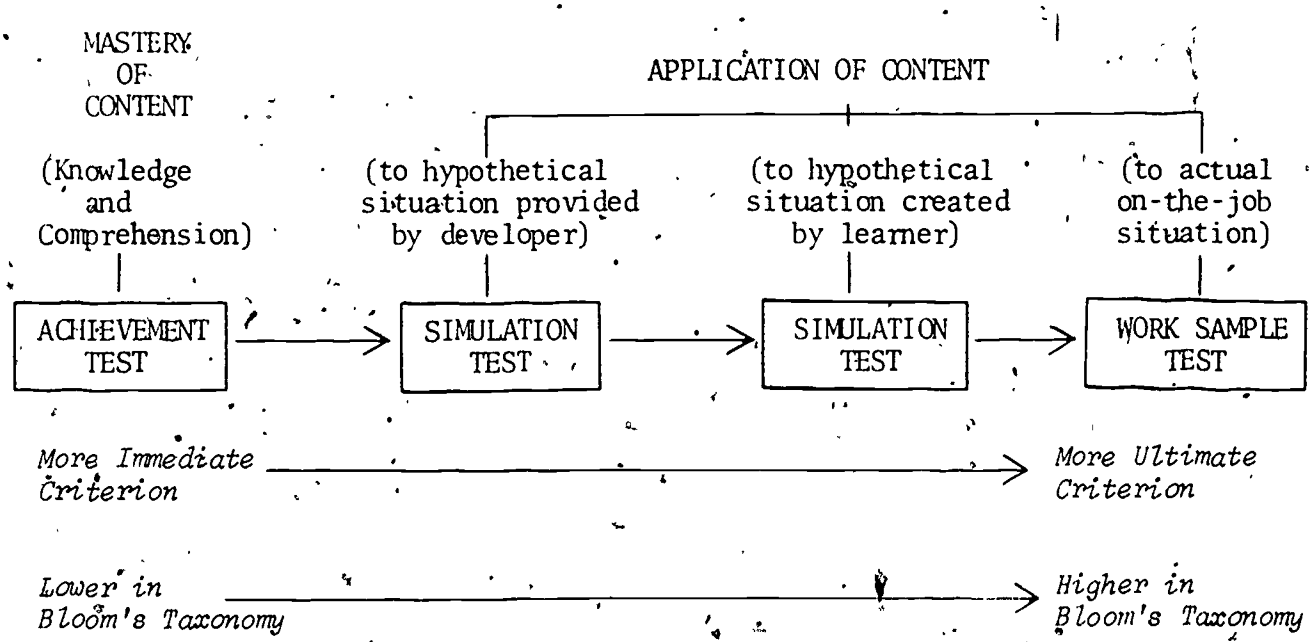
We might observe that the more narrow specification of sub-goals is merely the result of an apparent behavioral objectives approach followed by the administrative product evaluators. This statement, while true, describes rather than explains. Administrative products may have to use more specific subgoals because their content areas are less universally known. "Science achievement (in first grade)" is likely to elicit more common definition than "ability to manage an evaluation." What skills are needed to manage a curriculum evaluation? Can we assume that they are well-known and agreed-upon? Probably not; thus the evaluator of an administrative training product needs to subdivide the content area into discrete parts and then measure attainment of the parts. He should also offer a convincing argument that the user who can perform successfully on each of the sub-tasks has in fact performed successfully on the overall task: he should show that the sum of the parts equals a whole. That is, the administrative product evaluator may need to establish either judgmental or empirical validity for his measures, unlike the student product evaluator, whose measures may already have established validities. Furthermore, the administrative evaluator may need to show, in addition to the sum of the parts equalling a whole, that the whole is somehow worthwhile -- that this prespecified goal is desirable. Why is it important for school administrators to manage curriculum evaluation projects? Aren't evaluation specialists supposed to do this? However, it seems unlikely that a critic would ask, Why is it important for elementary students to learn science?

In other words, while product effectiveness was defined in all four cases as the match between goals and learner performance, the definition (and promotion) of these goals appears more difficult for the administrative product evaluator.

Criterion measures of effectiveness. Bloom (1956) describes cognitive objectives along a sequential taxonomy, beginning with knowledge and then followed by comprehension, application, analysis, synthesis, and evaluation. Both content "mastery" (i.e., knowledge and comprehension) and application are represented by the criterion measures used in the four evaluations described above. The science curriculum measured knowledge and comprehension via an achievement test. The proposal development product measured knowledge (but not comprehension) via an achievement test and measured application to a hypothetical situation provided by the developer via a simulation. The first curriculum evaluation product measured application to a hypothetical situation created by the learner via simulations. And the revised curriculum evaluation product measured application to an actual situation via work samples. We may conceptualize these measures of the criterion graphically, as in Figure 1:

FIGURE 1

Criterion Measures of Effectiveness


MASTERY OF CONTENT
APPLICATION OF CONTENT
(Knowledge and Comprehension)
(to hypothetical situation provided by developer)
(to hypothetical situation created by learner)
(to actual on-the-job situation)
ACHIEVEMENT TEST
SIMULATION TEST
SIMULATION TEST
WORK SAMPLE TEST
More Immediate Criterion
More Ultimate Criterion
Lower in Bloom's Taxonomy
Higher in Bloom's Taxonomy

In other words, each measure corresponds to a different point along the continuum. The dilemma of the administrative product evaluator is this: if he attempts to measure effectiveness via an achievement test, he is subject to the criticism that an administrative training product should produce more than "paper-and-pencil mastery" of cognitive content; on the other hand, if he attempts to measure effectiveness via a performance test (simulation or work sample) and results are negative, he will be unable to say whether poor performance data is due to learners' failure to master cognitive content, their failure to translate mastery to performance, or both. Another dilemma arises when choosing between simulation and work sample: while a simulation is more controllable, work samples may offer more realistic (and hence more valid) measures of performance.

Naturally, the way in which product goals are written may guide the evaluator toward the most appropriate measure. Is the administrator being trained to "master the principles of proposal development" or to "apply the principles of proposal development to an on-going project in his district?" Sometimes, however, product goals are so general that any of the four measures above could be used. The question is: Which one? Student curriculum evaluators are usually spared from this decision. We rarely ask a student to build a battery, just identify parts of a battery in a picture.

<u>User's pre-treatment level of ability</u>. The science evaluator administered identical forms as pre- and posttests of learner ability. The proposal development evaluator also administered identical forms as pre- and posttests. In neither evaluation of the curriculum evaluation product were users' pretreatment level of ability determined, apparently for two reasons: (a) the

criterion measures were highly idiosyncratic to the instructional materials, and it may have been unfair to pretest users unfamiliar with the terminology and organization presented in the materials; (b) as the focus was on group rather than individual performance, pre-measures of individual ability were irrelevant.

Let us examine each of these reasons for failure to obtain pre-measures.[2] That the measures were idosyncratic may be reflective of the product itself -- that it proposes terminology and organization different from those in popular use. However, it does not seem unfair to study whether learners have certain knowledge or skills, no matter what terms or procedures they use. The task of the evaluator is to develop a measure of criterion performance that is independent of the product (that is, free of its idosyncratic terminology and organization). The second problem -- how to measure group performance -- is more perplexing, especially if the measure employed is an on-the-job work sample. Since the group in training may not have worked together as a group before the training, any pre-training work samples collected may have been produced by a different set of individuals (i.e., a different "subject") in the same district.

Types of comparisons. In general, three types of comparisons may be useful in determining product effectiveness: (1) pre- vs. post-measures, (2) observed performance vs. desired performance, and (3) treatment group vs. comparison group.

[2] We say "failure" because these evaluations did not employ control groups either.

Comparisons between pre- and post-measures yield "gain scores" which cannot be considered treatment effects without benefit of a comparison group or other control. Similarly, discrepancies between observed and desired performance are merely descriptive without the experimental control provided by comparison groups. Thus, the identification of comparison groups is an assignment of paramount import to the evaluator who wishes to make inferences of cause-and-effect.

Note, though, that comparison groups were used in none of the evaluations of administrative training products. Why is this so?

When evaluating performance of school administrators, we often are less concerned with individual gains than with organization gains. We want to know, How much has the school district improved in its ability to manage a budget; train staff; evaluate programs; bring in new monies; etc.? Furthermore, we want to know whether the intervention (an administrative training product) has accounted for this improvement. Thus the focus is on organizational performance. So the comparison group must be composed of similar organizations -- but similar in what ways? On what organizational variables should we "match" school districts: pupil enrollment, location, organizational climate, organizational structure, personal characteristics of administrators? Without empirical evidence to show which variables moderate organizational performance on a given task, there is no guide for the administrative product evaluator to follow.

Sample size and comprehensiveness. The student curriculum evaluator generally has available to him a large student population using the new product under field test conditions. He may include all field test sites or sample from them, but either way the sample size is usually large enough to permit powerful statistical tests. Furthermore, the sample can be described in demographic terms (urban/rural, minority representation, SES, etc.) and can be shown to be representative of a larger population (i.e., the projected users of the product). A sample that is both large and representative may be called comprehensive.

On the other hand, the administrative product evaluator does not have available to him a field test sample of thousands, or even hundreds, or prospective users. A district may have 20,000 students, but it has only one superintendent. And it is often more difficult to persuade a superintendent or other district-level administrator to participate in an administrative product tryout than it would be to persuade him to volunteer a sizable portion of his 20,000 pupils for a student curriculum product tryout. District participation in a student tryout may yield a sample of 500; district participation in an administrative tryout may yield an n of 1, 2, or 3. Thus the job of the administrative evaluator can become overwhelming simply to execute so basic a task as identifying a sample of adequate size.

If the administrative product is to be used in workshop mode, it may be possible to involve 20 or 30 users in a single workshop. If, on the other hand, the product must be used on-site, a one-day tryout may involve but a few users. In such a case, multiple tryouts must be scheduled just to include as many as 20 field test participants. And if the organization rather than the individual is the unit of analysis, the effective n may only be 4, not 20.

Further, the determination of what is representative can be elusive. Which variables should be described? For instance, with the high rate of administrative mobility, does it make sense to label an individual as an urban administrator, when he may be a suburban administrator next year?

Acceptable criterion performance. When achievement test scores are used as criteria, as they often are in student curriculum evaluations, acceptable performance is a significantly higher mean score for the treatment group than for the non-treatment. But administrative training products can rarely employ an achievement score as an index of acquired skill or training in an executive function. Rather, simulation or work sample varieties of performance testing are more frequently appropriate. Scores on these types of performance tests may be heavily dependent on expert judgment, which is often highly variable across judges and across subjects. The low reliability and concomitant large error of measurement in such judgment scores hamper the decisiveness with which the evaluator can state that one group outperformed another.

Confounding variables. Both confounding and intervening variables can be reasonably controlled in student curriculum product evaluations through use of comparison groups that experience everything the treatment groups do -- except the treatment. With administrative products, an important source of confounding is the fact that administrators are being "trained" at all. The comparison group should therefore also be "trained," using a different training product or method. But, as mentioned earlier, it is often difficult enough to secure a treatment sample of school administrators, let alone a non-treatment sample who must also be trained. A second important source of confounding is the motivational level of administrators who desire a long-term working relationship with the development agency. Again, only if the non-treatment group is offered a similar opportunity to develop a working relationship with the development agency can the confounding be controlled.

Intervening variables. It may be reasonable to expect that comparison groups of third-graders are exposed to similar school experiences during the term. It is far less reasonable to expect that on-going professional experiences of school administrators are similar throughout the year. Some members of the field test sample may attend a conference, others may not. Some may be involved in heated teacher contract negotiations, others may not. Some may benefit from expanding budgets, others may be plagued by shrinking budgets. Each of these variables may affect criterion performance in a significant but unknown way.

Thus far, we have tried to point out some of the difficulties and inherent limitations in evaluating the effectiveness of administrative training products. Based on our review of three administrative training products and one student curriculum product, we have noted that:

1. Administrative evaluators must often specify product goals in almost behavioral terms, as there is seldom common understanding and definition of more global administrative goals.

2. While administrative evaluators may demonstrate product effectiveness via user mastery of content, such a measure may be inappropriate for products designed to train new skills.

3. Simulation and work sample varieties of performance tests, which may be more valid measures of administrative product effectiveness, often depend on expert judgments that are unreliable, creating large errors of measurement.

4. Work samples, which are more valid than simulations, are far more difficult to control.

5. Criterion measures are often idosyncratic to the training product, hampering measurement of both users' pre-treatment level of ability and non-treatment groups' level of ability.

6. The focus on organizational rather than individual performance makes pre- vs. post- and treatment vs. non-treatment comparisons difficult.

7. Large, representative samples of school districts (and/or school administrators) may be difficult to identify and involve in field testing.

8. Control of confounding and intervening variables is poor.

In light of these difficulties, it may be unreasonable to expect that administrative training products follow the same evaluation strategy as student curriculum products. It is hoped that there are forthcoming from the educational R&D community other strategies that promise to overcome some of these weaknesses in evaluating the effectiveness of administrative products.

A SUGGESTED PROCEDURE

The evaluation strategy suggested here is just that: a suggestion, not a prescription. If followed, it may overcome some of the limitations of current evaluation designs.

A logical starting point in product effectiveness evaluation may be to ask, "What kind of claim do the developers and/or evaluators wish to make regarding the effectiveness of the product?" Since the claim is dependent on the strategy used, the claim desired may have implications for the evaluation requirements -- e.g., the criterion measure of effectiveness, the sample size, the evaluation setting, and so forth.

For example, the evaluator of the curriculum evaluation product may present the developers with the following list of possible claims and ask them to rank the claims in order of preference:

a. This product is effective in teaching school administrators the principles of managing a curriculum evaluation project.

b. This product is effective in teaching school administrators how to apply the principles of managing a curriculum evaluation project to a hypothetical situation created by the developer.

c. This product is effective in teaching school administrators how to apply the principles of managing a curriculum evaluation project to a hypothetical situation in their own creation.

d. Under proper conditions, this product may be effective in guiding school administrators through certain activities in an ongoing curriculum evaluation project.

e. Under proper conditions, this product may be effective in guiding school administrators through a complete curriculum evaluation project.

Each claim suggests a different evaluation strategy. Claim (a) calls for an achievement test, Claims (b) and (c) a simulation test of performance, and Claims (d) and (e) a work sample test of performance. The design for Claim (a) could resemble that of the traditional student curriculum product evaluation. A design for Claims (b) and (c) might be as follows: randomly assign administrators from the same district to treatment and comparison groups. After training, both groups will be asked to perform the same simulated management activity. If individuals work collaborately on the activity, then there will be a single performance score for the treatment and for the comparison group. To employ analysis of variance, there probably should be a minimum of five such districts in this design, so that the organizational n equals 10 (5 treatment and 5 comparison). The use of administrators from the same district should control, to a large extent, confounding and intervening variables. It also permits reasonable comparisons between "equivalent" organizations.

A similar design for Claims (d) and (e) is possible, although it seems unlikely that two groups of administrators would actually work on the same management activity in a real situation. Therefore, in practice the requisite strategy for Claims (d) and (e) may comprise a work sample test of performance, a single group of specially-selected school districts, and an anecdotal history of how the product worked in each district. Since the latter design cannot be used to support causal relationships between treatment and effect, the resultant claim is necessarily equivocal ("Under *proper* conditions, the product *may be* effective...."). At best, the evaluator may be able to suggest what conditions seem to be proper for effective product use. Such equivocality should not be taken as a sign of low-calibre evaluation: rather, it shows that the developers have attempted to field test the product using a more ultimate criterion.

However, it may not be best to select a single, most desirable claim. The problem with choosing a single claim is that, because we might not know the exact relationship between content mastery and content application in managerial training, interpretation of the claim is difficult. Suppose we attempt to collect evaluation data to support Claim (b) and find that administrators are unable to demonstrate application of the principles taught. Is the product unsuccessful in teaching the principles themselves, or is it failing to help users translate the principles to a hypothetical application? We do not know, unless we have evaluation data on both content mastery and content application.

Therefore, an optimal evaluation procedure would be multi-stage, each stage focusing on one of the three criterion measures discussed earlier (achievement test, simulation, and work sample). An evaluation report that provides information on (1) how well users master content, (2) how well users apply the content to a hypothetical situation, and (3) what actually happens in the school district would clearly be a cut above those now offered.

Reference Notes

1. Evans, W. Broad scale report. EPMIS module 3: Proposal development instructional materials. Philadelphia: Research for Better Schools, 1974 (unpublished report).

2. Behrman, E. H. School evaluation kit: Limited test report. Philadelphia: Research for Better Schools, 1974 (unpublished report).

3. Behrman, E. H. Broad test plan for the Administrator's Handbook on Curriculum Evaluation. Philadelphia: Research for Better Schools, 1974 (unpublished report).

References

Bloom, B. S. (Ed.). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay, 1956.

Bloom, B. S., Hastings, J. T., and Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

Evans, W. J. The assessment of cognitive and affective outcomes of individualized science (RBS Publication RR-203). Philadelphia: Research for Better Schools, 1973.

Gagne, R. M. Curriculum research and the promotion of learning. In R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), Perspectives of curriculum evaluation, AERA Monograph 1. Chicago: Rand McNally, 1967.

Grobman, H. Evaluation activities of curriculum projects: A starting point, AERA Monograph 2. Chicago: Rand McNally, 1968.

Scriven, M. The methodology of evaluation. In R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), Perspectives of curriculum evaluation, AERA Monograph 1. Chicago: Rand McNally, 1967.

Tyler, R. W. Changing concepts of educational evaluation. In R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), Perspectives of curriculum evaluation, AERA Monograph 1. Chicago: Rand McNally, 1967.