UTILIZING EXPERTS' JUDGMENT IN THE PROCESS OF CURRICULUM EVALUATION

Arieh Lewy*

CSE Report No. 87
July 1973

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

*The author is Director of Evaluation, Curriculum Center, Ministry of Education, Israel, and Senior Lecturer in Education, Tel-Aviv University. Dr. Lewy spent the 1972-1973 academic year at the University of Pittsburgh and at the Center for the Study of Evaluation.

The collection and the analysis of empirical data with the intention of providing suggestions for the modification of curriculum material has been termed "formative evaluation" (Scriven, 1967). The most frequently used data, in such contexts, are responses to specially devised tests, and records of structured and unstructured class observations. The possibility and the limits of using experts' judgments for the sake of formative evaluation has not yet been systematically explored, though data of this type is frequently used in the process of curriculum evaluation.

Experts in curriculum evaluation have indicated the importance of judgmental data. Stake (1967) included "judgment" as a genuine part of a data matrix, which constitutes the basis for the inference about curriculum material. According to Fox (1971) the expert judgment uttered within the framework of the deliberations of the curriculum team (and refined and revised in face of opposing and/or supporting judgments of other experts) is the most valid evaluation strategy. Fenton suggested a series of questions which could be presented to experts, relying on "consensus among experts as the test of the validity of expert judgments" (Fenton, 1973).

The importance of judgmental data is indicated not only in theoretical papers and in models; practitioners also report the use of such data in actual work of curriculum evaluation (Flanagan & Jung 1970; Lewy, 1971). Therefore, it is rather surprising that only a few reports are available describing the actual utilization of judgmental data in curriculum evaluation (see, for instance, Peri, 1973), and that no attempt has been made to develop a methodological framework which may serve as a guideline for collection and analysis of data of this type.

The Place of Judgmental Data in Formative Curriculum Evaluation

Formative curriculum evaluation is not a one shot action; it is rather a sequence of activities which runs simultaneously with the curriculum development process from its very beginning. The evaluator, who is responsible for detecting weaknesses in the curriculum material, has to look for flaws as soon as they are detectable. Some flaws will emerge only after the curriculum is put to use in a number of schools, and these will remain undetected through the process of judgment. But experimental tryout can take place only when the material has already been developed to such a standard that school systems are willing to use it, while judgment of the material can be done at earlier stages of curriculum development. Thus, as soon as some portions of the curriculum material have been developed in first draft form, it is possible to critically review it. Experts may examine the preliminary draft of a single expisode, exercise, or other learning activity even before their sequential order in a course of studies has been determined. It is not implied that such a critical review alone is sufficient for evaluating the quality of the study material, but it is our contention that this method helps to detect flaws in the material at an early stage of the program development, when no other method can be used for this purpose.

The employment of judgment as an evaluation tool at the early stage of curriculum development is advantageous for several reasons. First, writers exhibit greater readiness to change elements of the program during the early stages of curriculum development than later after investing much work and energy in planning and in writing. Gradually writers become apologetic about their work and are inclined to justify what they have produced, rather than to accept criticism and re-write portions of the program. Secondly, there is

great economic advantage in using such an evaluation tool. An expert's judgment is relatively uncostly in terms of labor and time. Also, in most cases, it is easier to manipulate the time schedule of experts than the time schedule of students. Frequently a certain unit can be tried out only when students have reached a definite point in their study; the tryout of new study material is then restricted by the sequence of activities and other conditions in the would-be experimental classes. Thirdly, utilization of expert judgment may decrease the time needed for program development. This is especially important in educational systems where alternative study material is not available. In such cases one has to consider the loss caused to students by the fact that they have to use an obsolete program while waiting for publication of the new material.

Limits of Using Judgmental Data

The utilization of judgmental data for the sake of curriculum evaluation may well be subject to some limitations. There may be situations or circumstances where such data possess a high level of validity, while in other situations they may be less valid. There is a need for a series of studies which will help to specify these limitations and suggest conditions under which the utilization of such data is advisable. While there is a need to increase our knowledge concerning several aspects of the evaluation process, the present paper will be primarily addressed to two such possible limitations: the "what" and the "who" questions.

What to Judge

The judgment of curriculum material prior to its tryout in classes may have two major concerns: first, the fit of the material to a set of specifications, and second, the prediction of the responses or reactions which the

material will elicit.

Figure 1 represents four elements related to the process of planning and implementing learning activities. The left column elements represent the written curriculum and the right column elements the implemented curriculum. The connecting lines indicate areas of evaluation activities. The dotted lines and the inner quadrangle words represent activities which can be performed at pre-tryout stage; the solid lines together and the outer quadrangle words represent activities which are typically performed at the tryout stage. Four investigation areas are defined and described below.

Specifications and Outcomes. The relationship between specifications and outcomes is the major topic of most evaluation studies. The instruments commonly used in this context are achievement tests of different types. In Figure 1 a solid line appears between these two elements indicating that this type of information cannot be collected before the experimental tryout of the curriculum material.

Specifications and Planned Activities. These two elements of the scheme are connected with dotted lines, suggesting that the relationship between program specifications and planned activities can fully be studied at the pre-tryout stage through judgmental procedures, and results obtained at this stage need not be studied again at the tryout stage.

Planned and Actual Activities. It might be useful to design pre-tryout studies with the aim to predict the degree of the implementation of the planned program. If experts predict that the program will not be implemented and they suppose that the actual activities will substantially differ from the planned activities, then the developers will be in a position to modify the program in a way that will increase the probability of its full implementation.

WRITTEN CURRICULUM IMPLEMENTED CURRICULUM

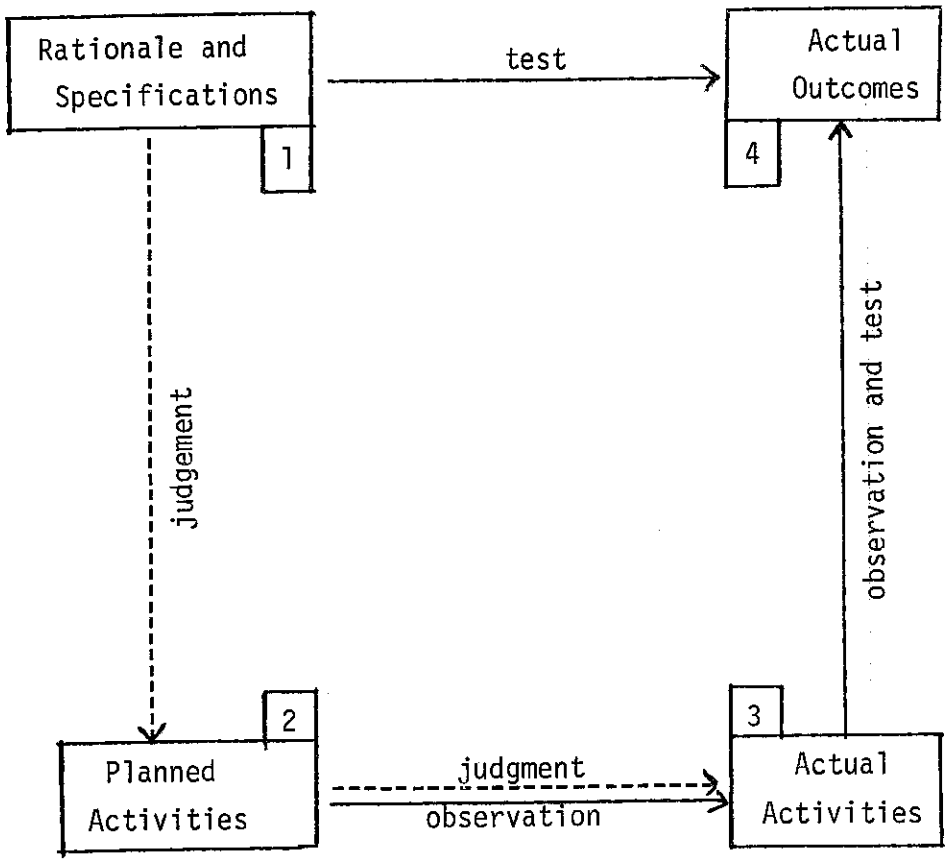

Figure 1. Planned and Actual Learning Activities

Legend

-------- pre-tryout
_____ tryout

In this area both predictive and empirical-observational studies are suggested. This is represented in Figure 1 by both a solid and a dotted line which connect these two elements in the scheme. One has to note, however, that in the normal course of curriculum development these two types of studies will relate to different sets of planned activities. The preliminary draft of a course will first be submitted to judgment, and on the basis of experts' opinion it will be modified. This modified version of the planned activities, and not the original one, will subsequently be put to experimental tryout and to an empirical implementation study.

Actual Activities and Outcomes. The relationship between the actual activities and outcomes should be empirically studied on the basis of observational and achievement data. These empirical studies are parallel to the logical investigation of the fit between specifications and planned learning activities. Studies of this type have been reported by Rosenshine (1971).

Thus, two areas are indicated where judgmental evaluation may be of value: one is the fit between specifications and planned activities; the other is the congruence between planned and actual activities.

Inventory of Questions

Several attempts have been made to prepare a list of questions which should be answered by experts with the aim to evaluate educational programs (Fenton, 1973). In Formative Curriculum Evaluation: A Manual of Procedures, Weiss, Edwards, and Dimitri (1971) compiled an extensive list of questions referring to various aspects of the curriculum, and suggested that evaluation experts should select questions which correspond to their interest. The present paper does not contain a detailed list of questions which may be presented to experts; it rather concentrates on indicating several aspects of the program which can be judged.

The Fit of Planned Activities to the Program Specifications

The questions listed here refer to specifications explicitly stated by the program writers and to those which are only implicit. The latter area includes such program specifications as presenting only scientifically valid facts and gramatically correct sentences. One may question the necessity of judging the scientific accuracy of curriculum material but a mere glance at some widely used textbooks in various subjects will convince the reader of the necessity for such judgment. Four major aspects of the fit and the considerations in each are listed below.

CONTENT	-- scientific accuracy
	-- significance of topics
	-- omission of issues of major importance
	-- fair representation of different views
PRESENTATION	-- correctness of language
	-- aesthetic value of illustrations
LEARNING APPARATUS	-- provision for individual differences
	-- clarity of cues
	-- variety of cues
	-- variety of learning activities
RELATEDNESS TO STATED OBJECTIVES	-- opportunity to learn what the program intends to teach

Prediction of Future Events

Judgments concerning the implication of the program and concerning the outcomes of the program in terms of students' achievements, teachers' reaction and community reaction constitute prediction of future events. Validity of such judgments can empirically be studied by later observing the process about which prediction has been made, while the validity of judgments of the previous type does not depend on events occuring in the future.

ACTUAL ACTIVITIES -- will the program be used properly

STUDENT INTEREST -- will students be interested in the material

DIFFICULTY LEVEL -- will the students encounter special difficulties in using the program -- does the text have an appropriate level of difficulty

TEACHERS' REACTION -- will teachers enjoy teaching the material
-- will teachers properly understand the material

COMMUNITY REACTION -- will people of the community find any of the material offensive, and interfere with the implementation of the program

Who Should Judge?

One should select experts in a way which assures maximum validity of their judgments. This implies that first one has to sort the questions according to the types of expertise needed to answer to them, and then one has to make a decision with regard to the individuals to be solicited. To judge the scientific accuracy of the material one will invite subject specialists. If within a single subject there are several alternative points of view, one should call upon a subject expert who identifies himself with the point of view which is reflected in the curriculum material. Also, if in a course in social studies conflicting opinions about a controversial social issue are presented, each of the different opinions should be examined by a person who is a prominent representative of that particular view, and each one should indicate whether the material properly represents that particular view.

The selection of experts for judging the fit between the program specifications and the planned learning activities should be made on the basis of logical considerations. As a rule of thumb it is suggested that experts should be consulted whose competency is recognized by the curriculum producers and by those

who will assume responsibility of making decisions regarding the use of the program. The selection of experts to predict events, to forsee actual learning experiences and outcomes, can be facilitated by systematic empirical research. Studies can be devised to explore the validity of predictions of experts of different types. By comparing a posteriori events with predicted events one may generalize about the ability of teachers, school administrators, school psychologists and educational psychologists to predict events of different types. Much can be learned from systematic studies of this type about proper preferences in selecting experts to predict future events. Studies of this type succeeded to provide useful techniques for improving the validity of prediction (Helmer, 1967).

The Value of Judgmental Data

The utilization of judgmental data is justified only if it detects flaws, the correction of which increases the quality of the curriculum material and improves the outcomes of learning. In other words, one should ask whether judgmental data are likely to yield valid conclusions. Three methods of examining the validity of judgmental data will be indicated below. They are listed in increasing order of strength to support the validity of conclusions derived from judgmental data.

Agreement Among Experts

One way to examine the validity of judgmental data is to quantify the degree of consensus obtained among experts with regard to different issues. Consensus among experts is usually considered as an indicator of reliability. If one uses the term validity in a sense of "well grounded, justifiable, true in terms of logistic system to which the inference belongs", at least with regards to some

questions, one may consider the consensus among experts as support of validity. Thus, for instance, the correctness of a statement, considering the present state of facts within the framework of a discipline, can be supported and proved by the consensus of competent experts.

It should be noted, however, that expert judgment should serve as a basis for the modification of curriculum material only if there is a considerable amount of consensus between experts with regard to certain issues. A mere majority of opinions should not necessarily require modification of the curriculum material. If a minority of competent experts have different opinions from those of the majority, the curriculum writers may be justified in following the opinion of the minority. Only an overwhelming majority of opinions should demand action.

The Shuffle Test

Another method of validating the conclusions based on judgmental data may be the employment of "the shuffle test." Using this approach, the curriculum team modifies the program according to suggestions emerging from the judgmental data; the two versions of the curriculum material, the original and the modified, are then presented to another team of experts who select the more appropriate one. It is assumed that if the modifications improved the quality of the program, the experts will prefer the modified version to the original one.

Preference for the revised version can serve as evidence of the contribution of expert judgment to the improvement of the curriculum material. Of course, the "shuffle test" can be relied on only if the modifications were satisfactorily carried out in line with the suggestions.

Experimental Validation

The most conclusive method of validating judgmental data involves experimentally validating the materials in operation. In order to conduct such an

experiment, the developers must maintain two versions of the material: one would be modified according to experts' judgment; the other would use only the internal review of curriculum. These two programs are then implemented in a random sample of classes; observations are made of the teaching process; and outcomes of the two types of study material are measured. Such a design can be employed only with regard to judgments related to future events. It can be used to examine the implementation of a proposed program but it cannot be used to examine the scientific accuracy and significance of the material presented in the curriculum material. Since this validation procedure is costly and time consuming it is not recommended for use in the normal course of curriculum development. Its use may be restricted to examining the validity of judgmental data in general.

Summary

Judgmental data are frequently used in social sciences, and their applicability to curriculum evaluation has been emphasized by experts. They can be produced at a low cost and at the early stages of curriculum evaluation when material is not yet ready for experimental tryout. Judgmental data may be used to evaluate the fit between the specification of the program and the planned or the written curriculum. It also can be used to predict the degree of implementation of the program. The validity of judgmental data can be examined in different ways including: consensus among experts, a shuffle test with regard to the original version of the curriculum material and the revised version, and experimental tryout of the original and revised program.

References

Fenton, Edwin. Criteria for judging curriculum materials before classrooms trial. Carnegie Melon University, 1973. mimeograph.

Flanagan, J., & Jung, S.M. An illustration: Evaluation, A Comprehensive Educational System. In Evaluative Research. Pittsburgh: American Institute for Research, 1970.

Fox, S. A practical image of the practical. Address delivered at the annual meeting of the American Educational Research Association Convention, New York, 1971.

Helmer, O. Social technology. New York: Basic Books, 1966.

Lewy, A. The practice of curriculum evaluation. Ministry of Education, Curriculum Center, Jerusalem, 1971.

Peri, M. Teachers make curriculum material. In A. Lewy (Ed.)., Studies in curriculum evaluation (in Hebrew with English summary). Curriculum Center, Ministry of Education, Jerusalem, in press.

Rosenshine, B. Teaching behavior and students. National Foundation for Educational Research in England and Wales, 1971.

Scriven, M. The methodology of evaluation. In R.W. Tyler, R.M. Gagne, & M. Scriven (Eds.), Perspectives of curriculum evaluation. AERA Monograph Series on Curriculum Education, No. 1. Chicago: Rand McNally, 1967.

Stake, R. The countenance of educational evaluation. Teachers College Record, 1967, 68(7), 523-540.

Weiss, J., Edwards, J., & Dimitri, O. Formative curriculum evaluation: A manual of procedures. Toronto, Ontario: Institute for Studies in Education, 1971.