THE EXPERIMENT
IN RESEARCH ON EVALUATION OF INSTRUCTION

M. C. Wittrock

CSEIP Working Paper No. 2, Dec., 1966
University of California, Los Angeles

The research and development reported herein was performed pursuant to a contract with the United States Department of Health, Education, and Welfare, Office of Education under the provisions of the Cooperative Research Program.

The Experiment in Research on Evaluation of Instruction

M. C. Wittrock

The study of educational evaluation, one of the most advanced areas of study within the field of education, has made great progress in ideas and methodology during the past thirty years. Most of this progress has been with the psychometric problems of selecting and placing individual students.

But the problems of evaluating instruction are essentially those of evaluating contingencies between instructional variables on the one hand and their multiple effects on the other hand. The thesis of this paper is that the evaluation of these contingencies is a problem that provides an excellent vehicle for developing knowledge about the process of evaluation. Further, it is maintained that the evaluation of contingencies should frequently involve the experiment as a preferred method.

In the literature of educational research, there are many experiments upon instruction, and there are many studies on evaluation. But there are few experiments upon the evaluation of instruction found in the literature, and very few of these experiments upon the process of evaluation per se, rather than upon instruction. The lack of interest in the experiment as a tool for evaluators to use in studying the process of evaluation has been accompanied by little progress

in the development of theory and methods for evaluation of instruction. Perhaps these two events are related by cause and effect more than by coincidence.

Since the work in the 1930's and 1940's by Tyler, Bloom, Troyer, Pace, and Lorge, there has been little advance in the methods of evaluation of instructional programs. We evaluate television today much as we evaluated motion pictures years ago. The new elementary and secondary curricula in the physical sciences are not being evaluated as thoroughly today as were the curriculum developments evaluated in the Progressive Education Association's Eight-Year-Study or the Cooperative College Study on General Education conducted years ago.

But new curricula and new instructional programs are being introduced into all levels of public schooling. In the last five years or so there has been a flurry of interest in the evaluation of instruction by men such as Cronbach, Stanley, Harris, Tyler, Lord, Stake, Scriven and Lumsdaine. There is use for new ideas and new methodological approaches to evaluate the <u>contingencies</u> between complicated instructional programs, including their important instructional variables, their administrative, social, and community contexts on the one hand, and their multiple effects on students on the other hand. The evaluation of contingencies between instruction and its effects raises new problems. Many of these problems have been thoughtfully discussed by the men cited

above. But I want to discuss problems of designing and conducting an experiment upon the study of evaluation of instruction, when that experiment is intended to develop knowledge about evaluation, but not necessarily knowledge about instruction. To bring the experiment to evaluators would seem to be equivalent to "bringing coals to Newcastle," but I shall run that risk. The study of evaluation of instruction can be improved if those interested in instruction exchange ideas with those interested in educational evaluation.

The Evaluation of Instructional Variables, Not Instructional Treatments

One of the first problems in the experimental study of evaluation of instruction for evaluation's sake is the identification of the important instructional variables within the instructional treatments...for example, a new mathematics program. A new mathematics program may vary from other programs in a variety of ways: in the number and variety of concepts presented in a program, the amount of reinforcement provided in the program, the order of presenting rules and problems, and the opportunity for the student to make overt responses. It makes sense to evaluate these variables rather than to evaluate the diffuse and less well defined whole treatment. Unfortunately, there are no taxonomies of educational stimuli or of instructional treatments which can be used to classify the differences among complex

curricula and teaching procedures by their meaningful instructional variables. Currently, an evaluator must turn to a specialist in learning theory to classify the treatments in terms of their instructional variables. In the experimental model of the evaluation of instruction, which is presented below, it is assumed that one should evaluate instructional variables, not complex instructional treatments.

The Evaluation of Interactions Between Instruction and Contexts, Not Only Primary Effects

The second problem in the experimental study of evaluation of instruction involves the deliberate complication of evaluation with measures of sociological context, of cost effectiveness, of teacher characteristics, and of student characteristics. Contextual variables must also be considered, The effects of instruction may depend to a great extent upon student characteristics such as social class and intelligence, A treatment effective for one intellectual ability may not be effective for another intellectual level. Experimental designs which do not include the important contextual variables are no longer adequate in the study of applied problems of evaluation of instruction. Obviously interactions between instructional variables and contextual variables will be discovered only when the experimental design allows such interactions to be evidenced. We should expect the effects of instruction to interact with at least some of the sociological, psychological, and administrative characteristics which comprise

the contexts of instruction. We should measure these interactions when they do occur.

The Evaluation of Effects of Instruction, Not Effectiveness of Instruction

Selection and measurement of the criteria of instruction is the third major problem area for the study of evaluation of instruction. Although it is largely accepted among researchers on evaluation that our dependent variables, i.e., our criteria, should be quantified with behavioral data, there is no consensus either about the form or the variety of these behavioral data.

There is the question of criterion-referenced data versus normative data. And there is also the issue of the use of gain scores versus post-test data in evaluating instruction. These, and other related issues in educational measurement should be left to specialists trained to handle them. These men should consider the possible utility for studying evaluation of descriptive measures of skewness, kurtosis, and of inferential statistics, such as chi-square, which compare whole distributions with each other. In evaluation of instruction, we should also be concerned with the effects of treatments upon the shape of distributions of student achievement. For example, does instruction change the distributions of subgroups of students from nearly normal ones to positively or negatively skewed ones? From certain instructional variables can we expect the distribution of achievement to become

more or less flattened or platykurtic? These types of measures would provide information not obtained by analysis of differences among means, the usual statistic in studies on instruction.

In addition to the above methodological issues, and other related issues such as whether or not test items should be written to discriminate maximally among individuals, or when we should use item sampling, etc., it seems plausible that we should measure the multiple effects of instruction, rather than merely to measure how well the objectives of instruction have been accomplished. The concept of the good teacher is superficial and oversimplified. The concept of effectiveness of an instructional program is also an oversimplified one, at least for a study designed to produce knowledge about methods of evaluation. Instructional treatments, especially when analyzed into their important instructional variables, can be expected to have multiple effects. There are few good or bad treatments in the simple-minded sense. For example, although a new mathematics program may well teach important concepts in mathematics to many students, one can still ask whether these students are learning to be attracted to mathematics. One can also ask how the conceptual styles of students are affected by the instructional variables. Does the instruction help to teach students a way of reflective or impulsive action about mathematical problems?

Another dimension of evaluation, also not commonly employed today in studies of evaluation, is the transfer and savings in future learning produced by instruction. For instance, to what extent does the instruction contribute to

learning more advanced concepts within the same discipline? To evaluate instructional programs, one should use multiple dependent variables, including at least measures of 1) conceptual style, 2) transfer, and 3) savings, to index the several effects of instructional innovations. We have been limited and uncreative in our conceptualization of the criteria of instruction. Many criteria should be sampled in research on evaluation.

A Research Design for the Model

The approach explained above can be useful in a study on evaluation. The true experiment involves random assignment to treatments. It is feasible with many instructional programs in many school contexts.

Let us assume we are interested in testing hypotheses about evaluation of instruction. A researcher on evaluation could introduce at random two different programmed mathematics curricula into several schools. These two instructional treatments would be described in terms of their instructional variables. The sociological and psychological characteristics of the students and the administrative characteristics of the schools would then be quantified. These two sets of variables, instructional and contextual, would comprise the input or independent variables in the experiment. The dependent variables, more comprehensive than are customarily used in studies on evaluation or studies on instruction, would complete the data needed to conduct this experiment upon evaluation.

At this point, one could ask how this procedure differs in principle from a conventional study of the evaluation of curriculum or instruction. Specifically, he could ask, how would the experiment be used to study evaluation rather than to study the differential effects of one math program compared with another? Obviously, the methodology of experimentation can be used to learn about either evaluation or instruction. But, primary objectives of the study of evaluation for evaluation's sake are to test ideas and hypotheses about models of evaluation, and to devise instruments and new procedures to enable one to discriminate the multiple differences among the instructional treatments. The hypotheses and the conclusions are all about evaluation.

If we assume that, like two children, no two instructional treatments are exactly alike, the experiment on evaluation leads to inferences about evaluation. For example, whether one instructional treatment is reliably more effective than the other is not a central issue in the evaluation study. If one finds that the instructional treatments are measurably and significantly different from each other, he could then conclude that his model and hypotheses about evaluation gained support, or that his measuring instruments were sophisticated enough to index the differences in learning produced by the treatment. He might conclude that he had identified and quantified some of the meaningful instructional and contextual variables which differed across the treatments. He would be using the experiment to study problems of evaluation. His

hypotheses, conclusions, and inferences are all about evaluation, not about instruction.

Summary

The experiment is a research design useful to the study of evaluation. In the evaluation of instructional programs, the primary problem is to find ways to relate outcomes to their probable causes...that is, to evaluate contingencies between instructional and contextual variables on the one hand, and multiple criteria on the other. The study of contingencies between instruction and changes in behavior is a fundamental problem in educational evaluation. The experiment has not often been used to investigate the phenomena of evaluation; but in the area of evaluation of instruction, the experiment has great promise.