COMMENTS ON PROFESSOR MESSICK'S PAPER ENTITLED "THE CRITERION PROBLEM IN THE EVALUATION OF INSTRUCTION: ASSESSING POSSIBLE, NOT JUST INTENDED OUTCOMES"

Leonard Cahen

Educational Testing Service

From the Proceedings of the

SYMPOSIUM ON PROBLEMS IN THE EVALUATION OF INSTRUCTION

University of California, Los Angeles
December, 1967

M. C. Wittrock, Chairman

Sponsored by the Center for the
Study of Evaluation

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.

CSE Report No. 24, May, 1969
University of California, Los Angeles

COMMENTS ON PROFESSOR MESSICK'S PAPER ENTITLED "THE CRITERION PROBLEM IN THE EVALUATION OF INSTRUCTION: ASSESSING POSSIBLE, NOT JUST INTENDED OUTCOMES"

Leonard Cahen

In his paper Samuel Messick covers many important aspects of evaluation, and especially emphasizes cognitive styles and affective reactions as they pertain to instructional research. In addition to the commonly assessed areas of pupil achievement, cognitive styles and affective reactions are suggested as further areas of possible assessment in evaluation studies. Among the major themes of Dr. Messick's paper are the need for assessing multiple dimensions of instructional outcomes, the importance of value judgments in instructional systems and their evaluation, and the role that individual differences in cognitive styles and information processing may play in future instructional research.

The idea of assessing the possible and not just the intended outcomes raises some important issues for the evaluator. At a first glance the term "intended" (the counter term "unintended") poses a difficulty. It takes a great deal of wisdom on the part of the evaluator to anticipate the unintended outcomes of instruction and to make the necessary plans for their assessment. In one sense the unintended outcomes may be conceived as unsought "side effects." In a hypothetical example a high school district adopted a new tenth grade science curriculum. The objectives of the

curriculum, among others, were to foster scientific thinking and to develop laboratory skills and an understanding of the scope of science. At the end of the year, the pupils performed satisfactorily on tests designed to measure these objectives. A negative "side effect," however, was seen in the fact that only a small proportion of these tenth grade pupils elected an eleventh grade science course the following year. The proportion of these students electing the eleventh grade science course was significantly smaller than the proportion of eleventh grade students taking science courses over the preceding years.

The curriculum builders and school administrators felt that there was a cause and effect relationship in the situation and decided that it was important to learn more about why the students generally failed to elect an eleventh grade science course. This negative "side effect" or unintended outcome was assessed by student interview.

A second form of the unintended outcome occurs when the curriculum developer explicitly attempts to develop a certain set of behaviors but not other behaviors. For this example let us assume that he is attempting to develop behaviors A, B, and C but not D. In this case the intended outcomes are A, B, and C where D becomes explicitly stated as an unintended outcome. An example might be found in one of the modern mathematics curricula developed in the

early 1960's. Behaviors A, B, and C might be represented by three sophisticated mathematics behaviors such as understanding of different number systems, the development of heuristics in problem solving, and an understanding of mathematical algorithms. Behavior D (the unintended outcome) might be represented by a traditional mathematical skill such as accuracy in routine computations or competence in translating Roman numerals. In the mathematics curriculum, the developer of the instructional program has in part exposed his value system.

The example of the mathematics curriculum represents a common problem that faced evaluators in the 1960's. The problem is manifested in the area of instrumentation where the instructional system or curriculum developer felt that standardized testing instruments failed to measure the dimensions he was interested in--A, B, and C (his intended outcomes)--but measured dimension D (his unintended outcome) with relative precision and validity. The mathematics curriculum suggested has led to considerable debate about the role of comparing outcomes across competing curricula or instructional systems when the competing systems have different intended outcomes.

Michael Scriven (1966) has introduced the terms formative and summative evaluation. Formative evaluation is the gathering of information in the early phases of developing a system of instruction. It is used for immediate feedback in modification of the materials.

Summative evaluation provides information to the potential consumers of the instructional product. However, as Scriven has pointed out, the distinction between the two terms is not always clear. If curriculum is to be an ongoing activity, a summative evaluation will serve as a first stage of a formative evaluation for the second wave of innovation. In the example of the mathematics curriculum developed above, the evaluator would be asked to provide information on dimension D as well as dimensions A, B, and C if the evaluation were summative.

The two examples of unintended outcomes are developed to show that an outcome may be an unsought side effect, unplanned by the innovator, or may reflect an a priori value judgment by the innovator to exclude certain dimensions from the instructional system.

Dr. Messick has urged evaluators to include psychological as well as achievement dimensions in the evaluative act. He has proposed that, in addition to assessing the face value components of achievement, instructional systems must also focus on processes and psychological variables as outcomes.

The issue of value judgments in evaluation cannot be overemphasized. Dr. Messick has pointed out that value judgments are made at many phases in the development and assessment of instructional systems. Judgments determine what the anticipated behavioral outcomes are, how they are to be reached, the components and constructs to be measured, and the selection of instruments or techniques

to measure or assess the components and constructs and, at a later stage, are used to reach decisions from the outcome data matrix. Too frequently value judgments, at least explicitly, are faced only at the decision-making stages, if at all.

Scriven (1966) has taken the position that the evaluator must play a key role in the incorporation of value judgments in the evaluative process. This is not easy task for the curriculum evaluator, and because he may not represent the specific discipline underlying the curriculum innovation he has felt that the judgmental processes must be left to the curriculum innovator who does represent the field. Robert Stake (1967) has hypothesized that the evaluator might have less access to data if he became identified with the judging of an instructional program. Stake also suggests the problem involved in judging the merit of a program from multivariate data where some of the outcomes are positive and supportive while other outcomes from the same program may reflect negative findings.

If we are to follow Scriven's suggestion that evaluators play active roles in the establishment and utilization of value judgments, we will probably have to give thought to the future sources of evaluators and careful thought as to their training. In addition, the need for identifying methods to analyze values reflected in a program or instructional system (and across competing instructional systems)

will hopefully be given more emphasis in evaluation enterprises of the future.

A proposal is made here that might complement methodologies in evaluating and assessing values in instructional research. The proposal states that outcomes at any stage of instruction can be assessed in terms of how well the instruction has prepared the students for future learning. An assumption is made here that learning is a continuous process and that school curricula will eventually reflect a continuity of experiences rather than inarticulated segments of curricula, i.e., elementary school math, junior high school math, etc. The success of an instructional program at any level could then be evaluated, in part, in terms of pupils' increased aptitudes for future learning.

I would now like to turn to the problem of utilizing individual difference data as elements in the process of placing groups of students in the most appropriate learning treatment. By most appropriate I mean the assignment of pupils to a learning situation or treatment where the pupil has the highest probability of maximum output or achievement. Dr. Messick has carried his suggestions past the initial stages of evaluation to the stage of implementation.

The model using the interaction of treatment or instruction and selected individual differences of learners has received a great deal of attention recently (Cronbach & Gleser, 1964; and Cronbach, 1966). While there is not always agreement about the results of such

an interaction model, the conceptualization does form interesting and explicit hypotheses and requires a major change in the application of quantitative strategies to education. It was not too many years ago that behavioral scientists hoped for non-significant statistical interactions in their analysis of factorial designs. Non-significant statistical findings at the interaction level allowed them (so they believed) to move on to the clear testing of major effects. Similarly, statistical textbooks frequently emphasized techniques for pooling the lower order interaction mean squares with the error mean square so that more stable error terms would be available for testing main effects. This technique of pooling reduced type two errors at the expense of potentially destroying the "nuisance" relationships displayed in interactions.

Dr. Messick has stated that interaction models may be useful in the examination of relationships between teacher and pupil characteristics on cognitive dimensions and in determining how these factors might interact to effect pupil learning. One may also wonder about the possible relationships between different organizations of the teaching act with pupil and teacher characteristics and how these would jointly effect pupil learning. Lastly, one may consider the relationship of individual differences on cognitive dimensions (teacher and pupil) and the structuring of the content of instruction. Might there be ways of organizing and presenting the content

of instruction so that it interacts with individual differences of pupils and teachers and teaching methods?

The use of individual difference interaction models will require concentrated efforts by evaluators to develop measures with minimal errors of measurement at the critical positions on the individual difference scales where decisions are made to assign pupils to learning experiences.

The technique of developing evaluation instruments for reliably measuring individual differences has recently given ground to the development of techniques to assess and evaluate group performance. Evaluation studies will need to determine both the important research questions and what mixture or combination of individual versus group assessments reflect the most appropriate techniques for answering the crucial questions underlying assessment and evaluation of a specific instructional system. The item or matrix sampling model developed by Frederic Lord (Lord & Novick, 1968) is a valuable technique for estimating group performance on many dimensions. Additional sampling combinations of items and subjects (successive matrix samplings) would provide a better estimation of the total covariance structure of the set of behaviors under investigation. However, as Dr. Messick points out, there are limitations and potential dangers in inferring performance of individuals from "averaged" or group assessments. This danger is probably more

severe in assessing personality dimensions than in assessing achievement output.

It becomes apparent to the evaluator that there is an almost infinite number of possible dimensions to assess and evaluate. The innovator-evaluator must decide which dimensions have the greatest potential for providing information for himself while also providing multi-dimensional outcome measures for the potential consumer. Value judgments again must play an important role. Explicit statements from the innovator-evaluator concerning priorities assigned to measures, and facts relating to which evaluative dimensions are not included in the study are crucial.

I would now like to consider a few problems that lie ahead in the utilization of individual difference measures in the cognitive style (non-achievement) areas in curriculum research. The study of individual differences in cognitive styles is in its infancy. Dr. Messick encourages use of longitudinal methods to study the long term interactions between achievement and such psychological processes as cognitive styles. It would be possible and highly desirable to readminister achievement and cognitive batteries over a long period of time and to study the covariance patterns over time within and between the achievement and cognitive domains. The processes underlying achievement and cognitive functions may both be changing, thus making the analyses themselves and the understanding of the analyses

very difficult. Witkin and his colleagues (Witkin, Goodenough, & Karp, 1967) have recently reported longitudinal and cross-sectional data on measures of cognitive style. More research of this nature will be needed if we are to utilize and understand cognitive styles and their potential for curriculum research.

Dr. Messick has called to our attention three other aspects related to the role of cognitive styles and curriculum research. He has told us that school experiences should foster an increase in the repertoire of styles for individuals rather than increase the competencies of an individual on a limited set of styles at the expense of other styles. The latter possibility is an inherent danger in the individual difference interaction model. It might be possible to structure an educational experience so that groups of students develop or increase their cognitive abilities along one dimension while failing to incorporate other styles into their repertoire. The educator must be very careful in structuring these experiences. If we take the dimension of tempo outlined by Jerome Kagan (1966), analytic versus impulsive styles, it is easy to let the semantics of analytic over impulsive determine what appears to be the obvious treatment--and desirable outcome. We must learn to know under what conditions it is favorable for a specific student to act analytically, under what conditions it is best for him to act impulsively, and then to determine a course of instruction that will foster both. It would

also be important to teach the student to decide when one style or the other is more appropriate or beneficial.

Dr. Messick has hypothesized that there may be some very important stages in the development of conceptual or cognitive styles, possibly in the very early years, prior to the organism being exposed to formal education. A great deal of research will undoubtedly be devoted to this area in the future.

My final point concerns the difficulty of administering non-achievement batteries in the evaluation of instructional programs. By non-achievement I refer to measures of personality, cognitive style, attitude, etc. The problem of invasion of privacy must be considered. In addition, how do students respond to tests not perceived as achievement measures? Students and school administrators will not see the relevance of non-achievement type tests to the evaluation of instructional outcomes.

We will need to convince ourselves first of the utility of individual differences such as cognitive style for instructional research, and then help the innovator to see the value of including these and other process variables in the instructional "package."

REFERENCES

Cronbach, L. J. How can instruction be adapted to individual differences? In R. M. Gagne (Ed.), Learning and individual differences. Columbus: Charles E. Merrill, Inc., Pp. 23-39.

Cronbach, L. J., & Gleser, G. C. Psychological tests and personnel decisions. (2nd ed.) Urbana: University of Illinois Press, 1965.

Kagan, J. Developmental studies in reflection and analysis. In A. H. Kidd, & J. L. Rivoire (Eds.), Perceptual development in children. New York: International University Press, 1966. Pp. 487-522.

Lord, F. M., & Novick, M. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Scriven, M. The methodology of evaluation. American educational research association monograph series on curriculum evaluation, No. 1. Chicago: Rand McNally and Co., 1967.

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

Witkin, H. A., Goodenough, D. R., & Karp, S. A. Stability of cognitive style from childhood to young adulthood. Journal of Personality and Social Psychology, 1967, 7, 291-300.