Course Evaluation on the Local Level

BURTON L. GROVER *

THE rather perplexing experience of trying to evaluate a new English course, which is described in this article, managed to provide both a method that is applicable to many school situations and a sharper awareness of the lack of meaningful curriculum evaluation in local schools. The method of evaluation turned out to be quite straightforward, but it became straightforward only after certain basic questions were clarified. The process of trying to clarify—and to separate—the key questions revealed some reasons why the performance of local schools in evaluating their own programs can best be characterized as dismal.

In evaluation it is easy to ask the wrong questions or to ask the right questions in such a way that very sophisticated tools are needed to avoid ambiguous and inconclusive results. Also, while a typical school has not found it possible to exist without books, buildings, teachers, and scheduling arrangements, such a school has found it possible to exist with only the most superficial evaluation of its effectiveness. Time and resources tend to go to meet the most obvious demands, and getting the school into operation is about as obvious a need as one can find. When curriculum evaluation is not necessary for continued operation, when its questions are ambiguous or complex or unasked and require complex tools to seek answers, and when there are limited resources left over from essential demands of keeping the school going, it is small wonder that educators retreat from the task.

The combination of traditional neglect of curriculum evaluation, the increase in new educational programs, and federal requirements for evaluation for federally supported projects has stimulated the growth of outside agencies ready to contract for evaluation services. It would be of some sociological surprise if these agencies, devoted as they are to the single enterprise of educational evaluation, did not greatly complicate the whole matter, duly or unduly, to the point where no one else dares venture into the field. A representative of one agency, in fact, argued that evaluation is so complex that it would be inappropriate and an oversimplification to create even a single taxonomy of evaluation designs at this time.[1]

Questions of Fact and of Value

The trend toward curriculum evaluation by specialized agencies does not mean that local schools should not or cannot perform their own evaluation. Certain evaluative approaches, while not meeting all ideal technical standards, can yet be both practical and meaningful. The approach used to evaluate part of a junior high English program was, to all appearances to those involved, a practical and meaningful one.

The basic assumptions of this particular approach were these: To make things man-

[1] Blaine R. Worthen. "Some Notions on a Taxonomy of Evaluation Designs." Symposium paper, American Educational Research Association annual meeting, Chicago, Illinois, February 1968.

* *Burton L. Grover, Associate Professor of Education, Western Washington State College, Bellingham*

ageable, questions of fact and questions of value should be not only clarified but clearly separated; questions of fact should be asked primarily in terms of actual pupil attainment of specific course objectives; questions of value should be asked primarily in terms of whether the objectives are worth bothering to attain. These are not necessarily new ideas—they closely reflect the thinking of programmed learning and instructional systems specialists—but they have not gained enough currency to be obvious in the face of a local curriculum evaluation task.

The development of these assumptions was based on two recent trends. One was the growing realization that pupil performance on general achievement measures is usually little affected by specific courses or specific learning experiences. Norton has argued persuasively that, given the high correlation between achievement and general intelligence measures, which are quite impervious to training effects, it is unlikely that a single course will greatly improve performance on general achievement.[2] Coleman Report data supported this contention by showing relatively low variation in pupil achievement attributable to between-school variation in instructional factors.[3] Consequently, the "no-significant-difference" result is usually obtained when one course is compared with another on general achievement measures. Scriven reached the same conclusion when he stated:

> The result of attempts to evaluate recent new curricula has been remarkably uniform; comparing students taking the old curriculum with students taking the new one, it usually appears that students using the new curriculum do rather better on the examinations designed for that curriculum and rather worse on those designed for the old curriculum, while students using the old curriculum perform in the opposite way. Certainly, there is a remarkable absence of striking improvements on the same criteria (with some exceptions, of which the most notable is the performance of students in studies of good programmed texts).[4]

A major implication of these arguments is that, unless very powerful measures and statistical designs are used, one should accept the likelihood that any curriculum will at most noticeably affect performance only on measures specifically attuned to the actual makeup of that curriculum and forget about proving the superiority of one course on general criteria equally applicable to other courses. Any course comparison must then be left to value judgments about which course has objectives that are more useful for later learning and life outside of school.

The acceptance of this likelihood coincided with a second trend—the current emphasis on specific behavioral objectives. Work with specific objectives has provided a literature describing tools needed for assessment of what a course actually does accomplish. Using specific behavioral objectives as a tool, curriculum developers and evaluators can become primarily concerned with how well a course accomplishes its own objectives without at the same time worrying about how it compares with other courses on general measures which are more closely related to intelligence and cultural background.

Once the expectation that a certain course will have a marked effect on general criteria is abandoned and once it is decided to use behavioral objectives, it is easier to proceed with a two-part evaluation based on the assumption mentioned earlier. The clear separation of value questions and empirical questions makes local course evaluation possible, and the use of behavioral objectives makes the separation of questions possible. The empirical or "how much" question asks how well a certain course enables pupils to attain specified behavioral objectives. The value or "should" question asks the worth of

[2] Daniel P. Norton. "The Potency of Educational Treatments." Paper presented at the First Pre-Session of the Wisconsin Educational Research Association, Manitowoc, Wisconsin, December 1967. Copy available from W.E.R.A., 126 Langdon Street, Madison, Wisconsin.

[3] James S. Coleman *et al. Equality of Educational Opportunity.* Washington, D.C.: U.S. Office of Education, U.S. Department of Health, Education, and Welfare, 1966. 737 pp.

[4] Michael Scriven. "The Methodology of Evaluation." *Perspectives of Curriculum Evaluation.* American Educational Research Association Monograph Series on Curriculum Evaluation. Chicago: Rand McNally & Company, 1967.

specific course objectives for pupils and for society. Both types of questions are implicit in any evaluation attempt, but they are difficult to handle unless clearly separated. In this recommended procedure, empirical questions are used to assess pupil learning only in terms of specific course criteria. Between-course comparisons are left to questions of value.

Despite the dangers of overemphasizing trivia and delimiting instruction to artificially narrow bounds, the use of specific behavioral objectives highlights the need for considered judgment of their worth. Behavioral objectives, then, are not only a tool that makes the empirical part of curriculum evaluation feasible, but their use also helps divide evaluation into two manageable parts rather than one confused and unmanageable whole.

Practical Evaluation

When applied to the evaluation of a new linguistics program taught as part of the English courses on the seventh- and eighth-grade levels in one school district located at Manitowoc, Wisconsin, the practicality of these evaluation ideas was demonstrated.

The development and initial implementation of the course followed from some basic decisions about what should be taught concerning the English language.[5] One basic decision was to introduce junior high pupils to some principles of linguistics based primarily on generative-transformational grammar. After one year of initial tryout and modification, the principal authors of the course, Thomas Swenson and Karl Hesse, spelled out 14 behavioral objectives for one of the general cognitive goals, namely, understanding of certain basic principles and terms of transformational grammar. An example was the following:

> When given sample sentences containing structures on modification or coordination, students should be able to construct sentences containing similar structures.

[5] Thomas Swenson, Karl Hesse, *et al.* "*Linguistic Grammar.*" Mimeographed paper. Manitowoc, Wisconsin: Cooperative Curriculum Development Center, 1967.

From the formulation of the behavioral objectives, it was a relatively simple matter to construct one test which measured attainment of the objectives. The result was a 100-item test with clusters of items, ranging in number from 2 to 20, for each objective. Pass-fail points in terms of items correct for each objective were arbitrarily set by the authors as criteria for pupil attainment of the objectives.

In a traditionally ideal evaluation, a criterion measure would not be used—and evaluation would be delayed—until there had been some prior investigation of the measure's reliability and validity. Preestablished empirical validity of a criterion measure is always desirable, but when items are written to correspond to specific objectives, the assumption of their validity is easier to accept without prior testing of the instrument. Gagné's discussion of the function of "criterion-referenced," as opposed to "norm-referenced" tests, pointing out the need to assess outcomes of learning rather than differences between learners, supported the initial inclination to proceed with the course evaluation on the basis of specially derived and untried measures rather than standardized measures of only partial relevance.[6] The extensive evaluation of the AAAS *Science—A Process Approach* curriculum also apparently followed this practice in its development and use of specifically-related, behaviorally-oriented "competency measures."[7]

In this evaluation of the English course, behavioral objectives were not always as rigorously devised and used as envisioned by their advocates in programmed learning and highly individualized projects. In these projects, it would be expected that all objectives would be rigorously behavioral, there would be a single test for each objective, and there would be extensive provisions for recycling students who did not attain objectives at the first post-testing. In this course, which had

[6] Robert M. Gagné. *The Conditions of Learning.* New York: Holt, Rinehart and Winston, Inc., 1965. pp. 257-60.

[7] Henry H. Walbesser and Heather Carter. "Some Methodological Considerations of Curriculum Evaluation Research." *Educational Leadership* 26 (1): 53-64; October 1968.

not yet worked out extensive provisions for individualization and the recycling of pupils, the measures of the 14 behavioral objectives were all wrapped up in one test given at the end of a time period allotted to the linguistics course. Not an ideal situation necessarily, but one which is more in line with typical organization of classroom instruction and one which could give a fairly clear picture of modifications necessary before the course would be taught the next year.

The general cognitive objectives of the course covered a wider area than that specified by the 14 behavioral objectives, such as application to both reading and writing, but most course activities were closest to the tested objectives. As a result, the course evaluation measured attainment of certain objectives in the one area while at the same time attempting to recognize that these specific objectives were not comprehensive enough to include all possible and hoped-for outcomes. The use of behavioral objectives does have dangers, clearly described by Atkin, of overemphasizing trivia, causing neglect of hard-to-measure outcomes, unnecessarily delimiting the educational experiences of a course, and in general causing instruction to "regress to the objectives."[8] The fact that the empirical course evaluation was limited to a subset of objectives emphasized the greater difficulty of managing a comprehensive evaluation. Nevertheless, the course evaluation proceeded on the belief that careful evaluation of part of a course is better than no evaluation or careless evaluation of all aspects.

Implementing a Design

Once the specific objectives and criterion measures had been formulated, the empirical part of the course evaluation became a matter of implementing a design that would answer three questions: "How many pupils who received the new program attained each objective at the conclusion of the program?" "Could the pupils have attained the objectives at the start of the program?" "Can gains in attainment of objectives be attributed to the program rather than other factors?"

In order to answer these questions in a typical school setting, an appropriate design will most commonly involve some arrangement of pretests and post-tests for both groups in the program and control groups not receiving the program. Unless experimental arrangements with random assignment of treatments to groups or individuals are feasible—and they often are not—the evaluation design will be of a quasi-experimental nature resembling one of those described by Campbell and Stanley.[9] These designs are not as efficient as true experimental ones, but they usually represent an acceptable blend of practicality and rigor.

With the assumptions specified previously, the first two questions above can be answered fairly well simply by pretesting and post-testing the pupils in the course with the criterion measure. With due regard for the fact that little relationship between gains and the course has been empirically proven, a limited design of this sort may be sufficient for some purposes. However, two changes in the design can make it stronger and provide a tentative answer to the third question, the one which examines the relationship between the curriculum and the attainment of objectives. One change is to withhold the pretest (but not the post-test) from a randomly selected half of the pupils in order to separate any gains associated with the program from those associated with the practice effect of testing. The other is to pretest and post-test similar groups that do not receive the program to see if they register gains because of factors other than the new course.

The evaluation of the linguistics course involved all of these features. Twelve classes, seven eighth-grade and five seventh-grade, were selected from those which received the program. In four of these classes, a random half of the pupils took the pretest; in the

[8] J. Myron Atkin. "Behavioral Objectives in Curriculum Design." *The Science Teacher* 35 (5): 27-30; May 1968.

[9] Donald T. Campbell and Julian C. Stanley. "Experimental and Quasi-Experimental Designs for Research on Teaching." In: N. L. Gage, editor. *Handbook of Research on Teaching.* Chicago: Rand McNally & Company, 1963.

remaining classes, all of the pupils took the pretest. Five classes, two seventh-grade and three eighth-grade, were selected from neighboring school systems for control purposes. These classes, which were receiving instruction quite dissimilar to the linguistics course in Manitowoc, took the same pretest and posttest at approximately the same time as the classes receiving the new program.

Examination of the Data

Examination of the data first of all ruled out the practice effect of testing as a factor relating to gains in attainment of objectives. Differences between pretested and unpretested pupils were negligible, slightly favoring pretested pupils in the control classes and slightly favoring unpretested pupils in the classes receiving the program. This interaction was not significant nor did it approach significance when analyzed by a two-way analysis of variance ($F = 1.08$), and can most easily be attributed to sampling error.

The data, as summarized in Table 1, clearly show that pupils in the program on the average attained more objectives at the end of the program than at the beginning, whereas negligible gains were registered by pupils in the control classes. With no random assignment of pupils to courses, random assignment of curricula to class sections, or random sampling of classes, it is doubtful whether any statistical test of significance of the data is appropriate. If such analysis were in order, an analysis of covariance considering each class (not each pupil) as an experimental unit, and the pretest mean as a covariate, would probably be the best procedure. Such an analysis, if classes were divided into four groups on a one-way layout according to grade level and type of program, would yield an F ratio of 44.80, a surprisingly high figure.

The figures in Table 1, together with a significance test if it is used, clearly establish that gains occurred in the classes taking the program and that they did not occur in the control classes. However, while it is comforting to know that the course produced some gains in attainment of acceptable objectives, curriculum evaluation is not of great value to the persons concerned if it goes no further. Once the fact has been fairly well established that some gains occurred because of the program, the important information is that which reveals the degree and kind of success attained by the pupils. The average pupil attained little more than half of the 14 objectives. This implied that, if all the objectives are still to be considered to be worthwhile, some changes in instruction should occur, probably in terms of lengthening the course or focusing the instruction more, to increase the success rate. A 50 percent success rate is not a clear signal to keep the course as it is. An objective-by-objective analysis was even more revealing because of the great variation of performance among objectives. The percentage of eighth-grade pupils achieving success at the end of the program ranged from 23 percent on one objective to 80 percent on another. The patterns of success, which objectives had high attainment and which had low attainment, varied widely among classes, suggesting that instruction among the classes varied in the degree of emphasis placed on each objective.

This type of information was needed for diagnosis of the curriculum and its instruction. The information about which objectives had low pupil attainment gave an empirical base for reconsideration of the in-

Grade Level	Number of Classes	Group Means and Gain Scores					
		Experimental			Control		
		Pre	Post	Gain	Pre	Post	Gain
7	5 2	2.96	6.49	3.53	3.36	3.78	.43
8	7 3	4.20	7.44	3.24	3.92	3.97	.05

Table 1. Average of Class Means of Objectives-Attained Scores

clusion of the subject matter represented by these objectives in the junior high program for the average pupil (or at least an impetus for further trial testing to see if the low scores reflected a measurement problem as much as a learning problem). The class-by-class, objective-by-objective breakdown indicated to individual teachers which specific areas needed greater attention in their own instruction.

The overall picture of the degree of gain registered by pupils implied a need for lengthening the program for all but the most able pupils, with perhaps more review and practice toward the end.

This, then, was the picture provided by the data from one course evaluation carried out on the local school level. One should note that the study measured pupil outcomes rather than less relevant and more process-oriented information such as teacher opinions of the course's success or the literary value of the course plans. The evaluation was made feasible by abandoning any attempt to prove the course's superiority over another course on general criteria and by concentrating on success of the course in terms of its own specific and unique criteria. Any comparative evaluation would be left to value comparisons of the criteria of different courses rather than pupil performance on common criteria (with appropriate consideration of whether the criteria are easier for the same pupils to attain in one course than in another).

It should also be noted that a person need not know how to use inferential statistics, such as analysis of variance procedures, to conduct such evaluation because it is questionable if such statistics are relevant for this kind of activity. Knowledge of how to figure averages and how to group data for tables and graphs should be sufficient.

The empirical part of the evaluation did not try to demonstrate the desirability of transformational grammar as a basis for the study of our language; that decision had been made earlier. The empirical part attempted to reveal only how well pupils succeeded in learning that which had already been deemed desirable to learn, as well as which subparts of the course needed most modification for a greater success rate. This two-part approach to curriculum evaluation may not be as exciting as a clear demonstration of the superiority of one curriculum on broad, common criteria, but such demonstrations are rarely found, and a fruitless search for them may in the end remove local schools completely from the evaluation of their own programs. □