Some Methodological Considerations of Curriculum Evaluation Research

HENRY H. WALBESSER*  
HEATHER CARTER

The demonstration and description of effectiveness are principal concerns of every curriculum evaluation. The evaluation designs constructed and used by the experimental curriculum developers to accomplish these purposes assume many and varied modes. Examples of this variability are observable in the published literature describing experimental science curriculum projects such as the American Association for the Advancement of Science—Commission on Science Education (1, 2), ESI (3), BSCS (4), PSSC (5, 6), and CBA (7, 8).

One approach is to allow the author of the instructional materials to view the units being taught and to evaluate their success or failure on the basis of his observations. This tactic is usually applied to small segments of a complete instructional program. The author must be able to make the personal contacts necessary. The procedure assumes a level of objectivity on behalf of the author which may be difficult to maintain with respect to the materials he has written.

Some curriculum developers recognize the need for product evaluation, but do not choose to conduct the assessment. Two practices are commonly pursued by these groups. One procedure is to acknowledge the responsibility to evaluate, but delegate it to another organization. For example, a professional testing organization may be employed to construct the assessment measures. When this strategy is adopted, the testing organization has the responsibility of examining the prepared curriculum materials and determining what learner outcomes are anticipated from the use of the materials. Once this task is accomplished the test makers complete their activity by devising what are thought to be appropriate test instruments. The testing program is administered. The results of the testing program are presumed to be an impartial evaluation of the effectiveness of the curriculum. Obviously the amount of communication established between the project staff and the test makers is a variable that directly affects the appropriateness of the developed measures.

A second posture assumed by some curriculum developers is to argue that a meaningful evaluation cannot be done. Those who embrace this position often argue that the available measurement tools are too crude to obtain the desired information. The usual consequence of adopting this position is to call for no evaluation.

Still another strategy of evaluation is available. The assessment of the effectiveness of a curriculum can be based upon behavioral objectives. This strategy calls for the curriculum developer to describe expected learner outcomes in terms of reliably observable behavior and to construct measures that assess the attainment of the specified behaviors. The appeal of this evaluation strategy...
rests with the intimate involvement of the authors of the instructional material in the construction of both the instructional objectives and their assessment.

**Methodological Observations on Curriculum Research**

There is no universal research design that can be applied uniformly to all curriculum evaluations. In fact, much of the "textbook" research methodology has only limited application to the design of an evaluation for an ongoing curriculum project. Many of the characteristics considered necessary in expositions of experimental design cannot be met in curriculum evaluations. For example, random selection of tryout centers from a national sample is impossible. Random assignment of subjects to classes is not possible, except for a few school systems. It is difficult, if not impossible, to maintain intact classes from year to year and yet many curriculum projects develop material intended for several school years.

In general, the present school environment can be characterized as one in which the opportunities for manipulation of experimental variables do not exist for all practical purposes. Needless to say, this places a major constraint on any curriculum evaluation design which is expected to be operational. One might conjecture that this is the nature of the institution, that schools have always been this way. Cremin (9), however, suggests this was not always true of schools in the United States. Robert Davis (10), at a recent conference on mathematics education research, argued most persuasively that if we are to be able to study schools experimentally it is necessary to retrieve the ability to manipulate variables within the schools. But such conditions do not currently prevail and this must also be recognized.

The cost is an important consideration in the design of an evaluation for a curriculum project. As Scriven (11) has indicated, with the development of an extended evaluation structure a curriculum project must commit the use of a considerable part of its budget to this part of its activity. Currently, evaluation services may have available as much as ten percent of the project's direct cost budget; few have more funds for this purpose and most have less. From the standpoint of cost analysis with respect to actual funds expended, experimental curriculum projects resemble development efforts much more than they do research efforts.

The curriculum researcher who designs an evaluation that ignores these experimental constraints is naive. It is the responsibility of those charged with the design and execution of curriculum evaluation to deal openly and rationally with these limitations, and to effect the most meaningful data collection possible.

There are certain general characteristics of curriculum assessment which should be satisfied by any evaluation design. First, the assessment should measure the aims of the curriculum as the curriculum developers envision these aims. Second, each assessment task should be acceptable to the curriculum developer as an evaluation of a particular curriculum aim. Acceptance of the assessment tasks is an endorsement of them as adequate measures of the aims. This acceptance also implies a willingness to employ the results obtained from these measures as accurate reflections of success or failure at meeting the aims of the curriculum. Third, the data are gathered in a manner that follows the accepted practices associated with the collection of any experimental evidence, including a sufficient description of procedures so that replication is possible.

Yet, what data are meaningful? As with the choice of experimental design, there are a number of answers to this question. The identification and description of the measurable unit to be used is an important consideration in any choice. Consider a curriculum to be a set of instructional materials made up of separate designated units where each unit has one or more aims. Whenever the curriculum developer produces such an instructional package, the priority assignment of evaluation is to supply data on the success or failure of the instructional units.

Observe the futility of comparison...
studies at this juncture of the evaluation. Until it is possible to decide whether the aims of each instructional unit are met by the set of experimental materials, there is little point to the study of other curricula. Because the immediate development problem concerns itself with meeting the aims specified for each particular instructional unit, the nature of the measurable unit is critical.

One Position on Evaluation

But how can the description of aims be made so as to assist in their evaluation? One answer is to construct all instructional aims as descriptions of observable behavior. That is, observe whether the subjects exposed to the instructional material are able to exhibit the behaviors described as the instructional aims in the objectives and on the basis of these observations construct a success-or-failure decision.

The immediate evaluation concern is the assessment of acquired behavior. With involvement being a theme of most, if not all, current experimental curriculum efforts, it is unlikely that an evaluation of behavior could be restricted to paper-and-pencil measures of performance and still directly measure the total array of expected behavioral acquisitions. In fact, the expectation is that most of the measures of behavior will require performance tasks for which paper-and-pencil items are unacceptable as adequate measures.

Assessment measures of behavior also raise interesting psychometric questions. For example, the behavior described by an objective is to be acquired by what percentage of subjects exposed to the instructional material? 50 percent? 75 percent? 100 percent? or how many? If the curriculum is intended for general education, say as an elementary science curriculum for the elementary grades, then 100 percent acquisition of the behaviors stated as objectives is desired. Suppose an assessment task is accepted by the curriculum developer as evaluating whether or not a subject has acquired the behavior described by an objective. The results of administering the task to a number of subjects are certainly not to be judged by the usual item analysis procedures. Furthermore, a collection of such tasks are not constructed so as to contain a range of item difficulties. The result the curriculum developer is working toward is acquisition of each of the stated behaviors by all subjects.

A Practical Example

What occurs in practice when such an evaluation strategy is adopted? The evaluation of Science—A Process Approach, an experimental elementary science curriculum developed by the Commission on Science Education of the American Association for the Advancement of Science, is one example. This project adopted the strategy of stating the instructional aims of the curriculum in terms of reliably observable behaviors.

The instructional material for each year of the seven years covered by the curriculum is called a part. The instructional units are called exercises. Each part consists of approximately 25 exercises. The exercises are written for the teacher. There are no textbooks written for the student. The objectives stated at the beginning of each exercise are descriptions of reliably observable behavior. The behavior objectives are descriptions of the performances all children should be able to exhibit after instruction. They are, in this sense, minimal objectives for each exercise. These behaviors are the ones being assessed in the immediate evaluation procedures.

The characteristics of the behavioral descriptions as they appear for this curriculum have been discussed by Gagné (12) and Walbesser (13). It may be helpful to describe the evaluation procedure which deals with the success or failure of individual exercises as a hypothesis. This hypothesis can be described in the following manner:

Research Hypothesis: An elementary science curriculum for grades K-6 can be developed so that at least 90 percent of the children tested and who were taught from the instructional materials will acquire at least 90 percent of the behaviors described as objectives.

The portion of the evaluation intended

October 1968
to test this hypothesis must concern itself with the acquisition of the behaviors described as objectives for each exercise. Certain other considerations may not be quite as clear and deserve some clarifying remarks. First, if it is the acquisition of the behavior which is to be assessed, then a direct recall task is not acceptable. Each assessment task requires a change of stimulus or change of context, if the subject’s response is to be acceptable as evidence of the learner’s acquisition of the behavior. Tasks which require only recall—use of the same stimulus in the same context as the instructional activity—are not acceptable as measures of behavioral acquisition.

A second consideration is the validity of the tasks. The authorship of assessment tasks best resides with the author of the instructional material. The author is in the best position to construct tasks which accurately reflect the competency he wishes to be attained. A measure of the validity of the assessment tasks constructed by an exercise author is obtained by presenting other experts from the same discipline with the description of objectives, the instructional materials, and a collection of possible assessment tasks. Some of the tasks are the ones constructed by the author of the exercise for a particular set of objectives and some of the tasks are unrelated to the stated objectives. Each member of the reviewing panel is asked (on an individual basis) to match objectives, instructional activities, and assessment tasks. If the panel agrees with the exercise author, then the tasks are accepted; if not, the tasks are rejected; new tasks constructed by the exercise author; and the reviewing procedure is repeated. The reviewing committee can be a committee of the development project such as the one charged with evaluation or a separate group convened explicitly for this review purpose.

When to administer these assessment tasks is an important consideration. The decision to be made is whether the administration should occur immediately after instruction has been completed for an exercise or to delay administration some predetermined period of time. The question then resolves itself into one of deciding whether or not these measures are to be concerned with the extinction of behavior as well as with its acquisition. In the test of the research hypothesis it is conjectured that the instructional materials do bring the population of learners to some stated level of attainment, a 90/90 level in this case. The assessment tasks represent the source of data used to test this hypothesis.

Some portion of the evaluation strategy must be designed to yield data concerning the acquisition of behaviors immediately after instruction for the objectives of each exercise. The assessment tasks which serve this immediate evaluation purpose are best administered immediately following completion of the exercise. The longitudinal study of how these separate behavioral acquisitions relate to one another is not a concern of the immediate evaluation.

Observe that the research hypothesis calls for something less than 100 percent of the children acquiring less than 100 percent of the described behaviors. Why? The answer is that 100 percent acquisition by all children was considered unrealistic for a curriculum which is to cover seven years of instructional materials. Individual differences in acquisition contributed to this decision. The mobility of subjects into classes without prior exposure to the curriculum also contributed to making the decision. It was conjectured that under these constraints the 90/90 level of attainment would provide an acceptable indicator of success.

The preparation of the instructional materials for *Science—A Process Approach* has followed this pattern: (a) development and revision in summer writing conferences beginning in 1963; (b) a year of field tryout; and (c) revision based upon assessment data and teacher feedback reports. The cycle established by (b) and (c) is repeated at least twice for each exercise.

The exercises were arranged in a sequence based upon learning hierarchies. One hierarchy was constructed for each process. Each learning hierarchy is named by a process label. These process labels include:
Classifying, Communicating, Inferring, Measuring, Using Numbers, Predicting and Using Space/Time Relationships. Each of the eight learning hierarchies is a developmental sequence of behavioral acquisitions. More complete discussions of learning hierarchies, their construction, validation, and interpretation have been written by Gagne (14, 15, 16). Specific references to the process learning hierarchies of Science—A Process Approach have been prepared by Walbesser (17), the Commission on Science Education Newsletter (18), and the Xerox Corporation (19).

The instructional procedures described in the exercises are written in a form which outlines the organizational and instructional strategies to be used by the classroom teacher. Fourteen tryout centers were involved in the third year of use in schools (20). The distribution of tryout classes at each grade level for the third tryout year was kindergarten, 16 classes; grade one, 22 classes; grade two, 22 classes; and grade three, 39 classes.

The data collected to test the first research hypothesis were obtained from the assessment measures administered to individual children by the teacher or by some other member of the professional staff in the tryout school. A minimum of three children were tested at the end of each exercise completed by a tryout teacher. No pressure was exerted on the tryout teachers to complete exercises on any sort of time schedule. The teachers paced their own instruction. The three children tested were randomly selected at the beginning of each month and no children were repeated until all children had been in a test group at least once. The assessment measures were the same for all tryout teachers. The assessment tasks were prepared by the project staff and the evaluation committee working in the writing session during the summer.

The assessment tasks (the instructions, description, and acceptable responses) for an exercise are called the competency measure for the exercise. Both the words and the materials to be used are specified. The collection of acceptable responses for each of the assessment tasks is also described for the teacher to help in making a judgment as to whether the performance of the child was acceptable. Other individuals besides the classroom teacher were asked to administer one or more of the exercise assessments as a check on the teacher’s interpretation of his own children’s performance.

In this examination of the stated research hypothesis, the intention is to consider only the data for the experimental exercises which have undergone an original draft and two revisions. This included only the first four parts of the experimental materials which are intended for the kindergarten and grades one, two, and three.

As a convenience in categorizing the competency measure data, four classifications of exercises were created. Group one is defined to be those exercises in which at least 90 percent of the children acquire at least 90 percent of the specified behaviors. Group two is defined to be those exercises in which at least 80 percent, but less than 90 percent, of the children acquire at least 90 percent of the specified behaviors. Group three are those exercises in which at least 70 percent, but less than 80 percent, of the children acquire at least 90 percent of the specified behaviors. Group four are those performance results in which less than 70 percent of the children acquire at least 90 percent of the specified behaviors.

Graphs One through Four summarize the results of the competency measure testing. Each of the four groups are shown.

Part One Findings

Graph One describes the observed success of the children, who were using the Part One instructional materials and who were sampled for testing, in acquiring the behaviors specified by the exercise objectives and measured by the individual competency measure tasks. Of the 24 exercises in the part, 16 or 66-2/3 percent are group one. Hence, two-thirds of the exercises in Part One support the research hypothesis. Of the remaining eight exercises, all refute the hypothesis, six report results at the group two
level of performance and two report results at the group three level of performance.

While the group two exercises do not support the hypothesis, there is a temptation to make the assertion that exercises at this level of performance could be changed from refutation to support by another careful revision. Consider that each of these parts has undergone two revisions already. The data collected on the competency measure tasks represent the assessment of behaviors acquired from instruction with the third edition exercises. To suppose that another revision will substantially affect the observed performance results, so that all or most of the non-group one exercises will become group one, is to go considerably beyond the data. However, it does seem reasonable to conjecture that revisions in the instructional activities of certain group two exercises might possibly affect performance changes from group two to group one. Considering the revisions as approximation procedures, such a conjecture for certain group two exercises is defensible.

Of course, to be acceptable these changes would need to occur without altering the statement of the exercise objectives. The most likely candidates for such changes are the group two exercises, especially those within two or three percent of being at the group one level. There is reason to be cautious about the potential of additional revisions, but also hopeful about their possible effect on this special subset of group two exercises. For those exercises reporting performance acquisitions further away from the group one level, there is no reason to suspect that additional revision will sufficiently alter the performance outcomes of these exercises after they have already undergone two careful revisions.

The two exercises with the lowest level of performance acquisition in Part One are both from the same process, Observing. The two group three exercises are Observing 2—Observing Temperature and Observing 6—Perception of Odor. The competency measure for Observing 2 was administered to 48 subjects. The lack of acquisition of two behaviors proves to be the principal contributor to the group three level of performance for this exercise: (a) distinguishing between the temperature in two different locations, and (b) distinguishing between the temperature at two different times. What can be done to alter this situation?

Additional revision is not an attractive alternative, since this exercise is now a product of two revisions which have not resulted in a 90/90 performance level product. One course of action is to postpone the instructional activities developed to aid in the acquisition of these behaviors. The postponement could be to a later time in the part or to a later part. An alternative is to eliminate these behaviors as components of the Observing process learning hierarchy. If the elimination strategy is adopted, then some consideration must first be given to assure that these behaviors are not needed as prerequisite to the acquisition of some subsequent behaviors in the Observing process or some other process.

One behavior appears to be the principal cause of the poor performance level obtained for Observing 6. The competency measure for this exercise was administered to 40 subjects. The one behavior most often not exhibited by these subjects was that of being able to construct an arrangement of objects by their odor. Some of the poor performance for this behavior may be explained by hypothesizing that some individuals are not able to distinguish between odors in a way similar to individuals who are considered
non-tasters. A rival explanation is that there are certain subordinate behaviors which need to be acquired first, and these have not been acquired by the subjects tested. If the physiological explanation is accepted, little could be accomplished by revision. If the rival explanation is accepted, then the revision would focus on the identification and description of these subordinate behaviors.

Part Two Findings

Graph Two describes the observed success of the children in acquiring the behaviors specified by the exercise objectives and measured by the individual competency measure tasks. The data are the test results of children who did use the Part Two instructional materials and who were sampled for testing. Of the 26 exercises in this part, 20 or 77 percent are group one and the remaining six or 23 percent are group two. Six group two exercises do refute the hypothesis. One of the six exercises is an alternate exercise at the end of the part, Classifying B, which does not contain objectives necessary to the Classifying learning hierarchy. The performance results of this exercise are important as evidence that this exercise does not achieve the acceptable level of success. Of the remaining five exercises, one is close enough to group one to suggest the possibility of a successful revision. The other four are not likely to be affected by additional revision, but are needed as subordinate behaviors in their respective learning hierarchies.

To retain the group one and group two exercises in their present form and at the same time attempt to design more effective exercises to substitute for the group two exercises, would seem to be the sensible strategy in light of the data and these conditions.

Part Three Findings

The competency measure results for Part Three exercise objectives are summarized in Graph Three. Of the 26 exercises in this part, four exercises report levels of acquisition which support the research hypothesis. A majority of the exercises, 16, report acquisition results at the group two level. There are three exercises at each of the group three and group four levels. That is, 22 exercises or 85 percent refute the research hypothesis. Again it might be argued that some of the group two exercises could be improved by a revision.

Further analysis reveals interesting position patterns in the part for the group three exercises. Two of the group three exercises were positioned at the end of the part and consequently were taught near the end of the academic year. Not all of the tryout classes reached that portion of the instructional sequence. Therefore, the sample of competency measure results is quite small in comparison to the earlier exercises of the part. The poor performance results may be an
outcome of this limited sample. The one exercise classified as group three which was not at the end of Part Three was Observing 17—Bacteria. The performance data obtained from the competency measure testing identify two behaviors as the principal contributors to this low performance. One, the children tested were unable, for the most part, to describe methods for showing that bacteria could be found almost any place. Two, the children were unable to demonstrate that it is possible to identify and describe small things by observing the organisms in clusters.

Using Number Relationships 7—Multiplication, Measuring 8—Are Pictures Always Life Size, and Observing 18—Five Plants were the three exercises at the group four level. In the Using Numbers exercise, the children tested who contributed low scores were unable to construct the product of number pairs using a repeated addition algorithm or using a number line. Two behaviors were not acquired by those subjects who contributed low scores to the Measuring 8 results: one, the children were unable to identify and name the linear scale on a map; and two, they were unable to state the actual size of an animal given an illustration of it and the scale used. The results of the competency measure assessment for “Five Plants” identify the competencies of being able to construct an experiment to test a hypothesis and being able to identify the variable to be controlled in an experiment as the two behaviors not acquired by the subjects tested.

**Part Four Findings**

Graph Four displays the results of testing for acquired behaviors stated as objectives in the twenty-six exercises of Part Four. Fifty percent or 13 exercises report group one results, twenty-seven percent or seven exercises, report as group two, four exercises are group three and two exercises are group four. The two group four exercises are Classifying 7—Classification Techniques—Coding Systems and Using Number Relationships 11—Metersticks, Money, and Decimals. The construction of a code that could be used as a multilevel classification scheme and the identification of the basis on which a classification code could be constructed are the two behaviors not acquired by the subjects tested. The Using Numbers process exercise attempts to help the learner acquire the competency of being able to describe metric measurement and U. S. Currency in terms of decimal numerals. The data suggest this exercise does not succeed in accomplishing this objective.

The data for Inferring 5—Tracks and Traces, a group three exercise, indicate that the children tested were unable to distinguish between an inference and an observation. This same population of subjects was also found to be unable to identify fractions from the largest to the smallest. These data came from the assessment measure results of Using Numbers 10—Fractions, also a group three exercise. The third group three exercise in this part was Using Space/Time Relationships 17—Revolutions and Angular Velocity. The low performance results for this exercise were obtained for subjects who were unable to identify and name the circumference of a circle or use the circumference to calculate the linear speed of a wheel.

**Observations on the Findings**

The findings indicate that the level of acquisition stated in the research hypothesis was not attained for all exercises in Parts One
through Four. That is, ninety percent of the subjects tested had not acquired ninety percent of the behaviors specified as objectives for all exercises in Parts One through Four. Certain parts fared better than the others. The research hypothesis was supported for 53 exercises from a total of 102 exercises, or the research hypothesis was supported slightly better than 50 percent of the time. Recall too that these data were obtained on behavioral acquisitions for exercises which had undergone two years of field testing and two revisions.

Why is it that the objectives are not met (behaviors not acquired at the desired level) with the 90/90 success level? What possible explanations are there for such results? An explanation is that the instructional activities failed. Perhaps the scientists and teachers who authored or revised the exercises were wed to a single position and did not alter the exercises. Examination of the various editions does not seem to support this explanation, beyond accounting for a few of the findings. The competency measures may not assess the behaviors described as exercise objectives. In a certain few instances this may be true, but it does not generally hold. In fact, the editorial step of matching tasks to the objectives reduces the likelihood of such a circumstance to near zero.

Another possible explanation is that the instructional activities do not correspond to the objectives or the assessment tasks. In designing and revising the competency measure tasks, the context or the stimulus was changed from that used in the instructional activities, but the objectives of the tasks and the exercise activities were checked to see that they were the same. Test anxiety might be proposed as an explanation for some of the lack of expected attainment, but this effect should have been encountered for most exercises and a differential effect on half of the exercises is not to be expected.

Another rival explanation for these results appeals to the assumption guiding the construction of the sequence of exercises. The instructional program (exercises, equipment, and competency measures) was developed to be in consonance with an assumption about learning. The assumption or hypothesis is this—given a terminal behavior to be acquired, there exists a set of prerequisite behaviors which must first be acquired by the learner. Because of this assumption, an argument might be made that for exercises whose test data lead to rejection of the research hypothesis the data were contributed by subjects who had not first acquired the necessary prerequisite behaviors—behaviors acquired from previous exercises in the learning hierarchy. Furthermore, the phenomenon can occur because of mixing the data derived from assessing subjects who have had one or two previous years' exposure to the curriculum with data derived from subjects who have had no previous exposure. Such an explanation obviously does not account for the eight exercises in Part One which do not support the research hypothesis, but it may explain the results observed for Parts Two, Three, and Four.

Findings in Special Student Categories

For the subjects exposed to Parts Two, Three, and Four of the curriculum, there did exist a mixing of subjects by previous exposure. For example, the subjects exposed to Part Two are in two categories, those who have been exposed to Part One the previous year and those who have not been exposed to Part One. For the population of subjects exposed to Parts Three or Four, there are
three categories, those who have had exposure to the curriculum in two previous years, those who have been exposed to the curriculum in one previous year, and those who have had no previous exposure. The learning assumption relating terminal behaviors to subordinate behaviors suggests several related research hypotheses.

Before stating these hypotheses, it is helpful to describe certain conditions of subject exposure to the Science—A Process Approach curriculum. Table One (see page 61) displays the various conditions of exposure for each part.

Research hypotheses are constructed from the categories shown in Table One. The Part Two research hypothesis is:

Research Hypothesis: The behavioral acquisition of Part Two subjects Category 1 is greater than the behavioral acquisition of Part Two subjects Category 2.

If this hypothesis is supported, there should be an observable difference in the level of behavioral acquisition for the two learner classifications.

Graph Five displays the data for the two student categories within Part Two. The horizontal axis names each of the four behavioral acquisition groups and the two student categories within each group. The vertical axis names the percentage of exercises within each group for the subjects tested. Eighty-eight percent of the exercises are group one for the second year students tested, while 50 percent of the exercises are group one for the first year students tested. These data offer support for the second research hypothesis for Part Two.

Recall that the second year students represent the desired length of prior exposure to the curriculum, before instruction with Part Two. These observations coupled with these second year student data suggest that Part Two does approximate the desired 90/90 level specified in the research hypothesis for that part.

Graph Six shows the data collected for the three student categories within Part Three. The horizontal and vertical axes are the same as those used for Graph Five with one exception. Three student categories, first, second, and third year, are named on the horizontal axis within each group. The research hypothesis for Part Three is:

Research Hypothesis: The behavioral acquisition of Part Three subjects—3 is greater than the behavioral acquisition of Part Three subjects—4 or 4' which is greater than the behavioral acquisition of Part Three subjects—5.

The third year students tested do show that 58 percent of the Part Three exercises are at the group one level. However, the predicted ordering of second year student results being greater than first year student results is not supported by the data. The
third year students, it should be remembered, represent the desired length of prior exposure to the curriculum, before instruction with Part Three. The acquisition data synthesized in this form do not support the Part Three research hypothesis, although there is evidence of a cumulative effect for third year students.

The Part Four research hypothesis is:

Research Hypothesis: The behavioral acquisition of Part Four subjects—6 is greater than the behavioral acquisition of Part Four subjects—7 or 7© which is greater than the behavioral acquisition of Part Four subjects—8.

The data collected for the three student categories within Part Four are shown on Graph Seven.

![Graph Seven: Percentage of Exercises](image)

The third year students tested do display the greatest percentage, 64 percent of the Part Four exercises at the group one level. The research hypothesis related to order is supported by the Part Four data.

In summary, the first research hypothesis raised the question of whether it is possible to produce a set of instructional materials for which 90 percent of the students acquire 90 percent of the behaviors stated as objectives. The data based upon the combination of all student categories of exposure to the Science—A Process Approach curriculum do not support the research hypothesis for any of the four parts. The Part Two results come closest to offering support for the hypothesis.

Three explanations of these results seem most plausible. One explanation is that the instructional materials have not been revised sufficiently. However, if this reason is entertained, it should also be recalled that these instructional materials have gone through an original draft and two revisions. Under the circumstances, the possibility of additional revisions achieving the desired results is questionable. A second explanation is that the 90/90 criterion is too high an expectation. However, if a curriculum is to be designed for a general education audience, then the 90/90 expectation for a successful program seems to be a reasonable one. Such a level of expectation admits that 10 percent of the student population would not be expected to meet the minimal acquisition level. The reported data suggest that the 10 percent figure is not large enough.

A third plausible explanation for the observed level of performance in Parts Two, Three, and Four is based upon the observation that the behaviors for all exercises have been sequenced into eight learning hierarchies, one for each science process that has been identified. Since these are validated learning hierarchies, there should be an observable learning advantage for those students who have been exposed to the previous parts. Three research hypotheses were examined in relation to the predicted ordering—one for each of the Parts Two, Three, and Four.

Support was found for the hypotheses associated with Parts Two and Four. A cumulative learning effect was observed in Parts Two, Three, and Four. The first research hypothesis (that related to reaching the 90/90 level), however, was not supported by the data organized in terms of student exposure. The percentage acquisition increased for the students with the appropriate amount of exposure to the curriculum, but still did not reach the desired 90/90 level for all exercises. If the desire is to construct a set of instructional materials appropriate for a general education audience, then the data organized by student exposure suggest that the
minimal, acceptable level of success may need to be lowered from the 90/90 criterion.

The reported data also raise interesting questions about the behaviors that can be expected to meet the 90/90 criterion. For example, what particular identifying, naming, demonstrating, constructing, and describing behaviors are acquired by at least 90 percent of the learner population? Are there any generalizations which can be constructed from these data about acquisition? Does the context of the assessment task affect the indicated measure of acquisition? What instructional conditions affect the acquisition of these various classes of performance?

The methodological questions raised by the current curriculum development activities of ongoing projects are not resolved by this discussion. However, some aspects of the problem as it relates to research have been raised and one set of assumptions described. The behavioral objectives and learning hierarchies strategy for curriculum research merits some consideration. Alternative strategies must also be heard and considered. The hope of this presentation is that other curriculum research tactics will be described and their application illustrated.

References


20. Austin, Texas; Baltimore, Maryland; Glencoe, Illinois; Ithaca, New York; Kern County, California; Monmouth, Illinois; Oshkosh, Wisconsin; Overland Park, Kansas; Peiham, New York; Philadelphia, Pennsylvania; Portland, Oregon; Seattle, Washington; Tallahassee, Florida; Tucson, Arizona; University of Chicago Laboratory School.
Copyright © 1968 by the Association for Supervision and Curriculum Development. All rights reserved.