DOCUMENT RESUME

ED 049 271 TM 000 423

AUTHOR	Meese, M. Kathryr
TITLE	A Model for Assessing Complex Educational Outcomes.
PUB DATE	Feb 71
NOTE	28p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Classroom Observation Techniques, *Curriculum Evaluation, Discovery Processes, *Educational Objectives, *Evaluation Techniques, Formative Evaluation, Measurement Techniques, *Models, Student Evaluation, Summative Evaluation, *Test Construction, Test Reliability, Test Validity

ABSTRACT

A general model was built for assessing important complex educational goals. Literature on performance tests, process goals, and verbal protocols were critically examined. Advantages of these methodologies were incorporated into the model which was tested in an individualized mathematics inquiry laboratory which used whole-task approach. Five process goals were chosen to represent mathematics inquiry. Task analyses and observations determined criterion behaviors for each. Criteria were outlined for test items administered to children who were also observed in class. Additional validation studies are needed but some behaviors of this educationally important complex goal appeared to be tapped. Appendices include the detailed and categorized process goal of analysis and planning for the task, the problem situation itself, scoring key, a sample verbal protocol, the coding scheme used in observing the child's approach to the task, and a classification of strategies used. (Author/TA)

A MODEL FOR ASSESSING COMPLEX EDUCATIONAL OUTCOMES

M. Kathryn Meese
University of Pittsburgh

INTRODUCTION

After three years of research, the Commission on Tests appointed by the College Entrance Examination Board (1970) has recommended that the Scholastic Aptitude Tests of verbal and mathematical ability be eliminated as tools for making decisions about college entrance. In their place, the Commission proposed a "flexible assortment" of tests of complex educational outcomes, such as commitment to social responsibilities, sensitivity, adaptability in novel situations, and artistic talent. This dissatisfaction with tests that sample relatively simple kinds of learning has been a recurrent theme in education, but few tests of these complex kinds of learning have materialized. As a result, we find educators paying homage to these complex variables, while assessing them impressionistically or not at all, since appropriate evaluation instruments and procedures are unavailable. Time and time again, we find that those involved in innovative educational programs report dramatic changes in their schools while the data from standardized tests of skill learnings reveal no significant differences between experimental and control groups. This suggests that either educational programs are having no impact or that tests are not detecting the changes; in many cases the latter conclusion seems more probable.

Much of the evidence we now have about complex educational goals has come from factor analytic studies such as those of Thurstone, Cattell, and Guilford. Unfortunately, it is not clear how this information can be used to develop measures to detect the ways in which learning of such complex goals is influenced by instruction. Because those who have studied outcomes such as creativity, problem solving, and thinking have failed to analyze them into behavioral components, we do not know which aspects of behavior should be selected for evaluating instructional programs designed to promote these kinds of goals. In discussing this issue, Glaser (1967) has cautioned that "if, indeed, complex reasoning and open-endedness are desirable aspects of human behavior, then this needs to be a recognized and measurable goal. Overly general, non-performance based objectives may force us to settle for what can be easily expressed and measured [p.2]".

In this paper I shall first present a general model for developing instruments to assess students' progress in curricula designed to promote complex educational outcomes. I shall then describe an application of the model to a test designed to assess outcomes of an individualized mathematics inquiry lab. Briefly I shall consider some of the ways in which the development of such a test affected the curriculum work.

THE MODEL

The general model proposed for assessing complex educational outcomes is a composite of three heretofore independent methodologies: the performance test, the focus on process goals, and the "thinking aloud" technique.

Let us first consider the performance test. Unlike the paper-and-pencil test, a performance test requires that the examinee carry out some activity in a standardized situation. His verbal and nonverbal behaviors are evaluated either by an objective performance score or by observing the way in which he responds. The picture completion, picture arrangement, block design, object assembly, mazes, and coding subtests of the Wechsler Intelligence Scale for Children (Wechsler, 1949), are excellent examples of performance tests using manipulative materials. The performance test permits a fair comparison of individuals because each has the same opportunity to perform. Cronbach (1960) reports that success on performance tests appears to depend less on habit and more on the ability to attack a new problem. Some limitations of performance tests are that items require a longer time to complete than those on paper-and-pencil tests, and that within the same time period one samples a smaller set of behaviors.

A second component of the model is that it focuses on the processes and strategies of complex goals rather than just the end products. Several studies are available in which data have been obtained on the dynamics of such outcomes.

An early attempt in this direction was made by the Eight Year Study sponsored by the Progressive Education Association (Smith and Tyler, 1942). Diagnostic instruments were built to measure three processes; the ability to infer generalizations from specific data, the ability to apply known principles in explaining new situations, and critical thinking per se. This work has served as a benchmark for many developers of tests of process goals.

The Balance Problems Test was designed by Cross and Gaier (1955) to assess the processes involved in effective utilization of facts and principles. The student was presented with six sets of problems of increasing difficulty. Principles and facts needed to solve the problems were hidden under tabs which the student could uncover. The test was used with some success to predict mathematics grades of high school students. The tab format was also used in the X-35 Test of Problem Solving in Science (Butts, 1964). This test measured four processes: (1) early formation of hypotheses; (2) specific experimentation with relevant variables versus random guessing; (3) introduction of controls to test the validity of hypotheses; and (4) attempts to verify hypotheses. College-level students were presented with two problem situations in science. A series of relevant, redundant, and irrelevant data were hidden under tabs. The data chosen by each student and the order in which they were chosen were rated by three judges (with high agreement) against the four processes defined above.

The tab format used in these two tests reveals what information the students have seen, and scoring the tests is not difficult. However, a serious limitation of this format is that one does not know if the students actually used all the information or how they used it. In addition, the criteria used to judge various patterns of scores are not explicit; and no attempt has been made by the authors to discover specific behaviors which were related to the stated goals. Furthermore, since the subjects of these studies are adolescents, one does not know how well the tab format would work with younger children.

Several instruments have emerged more recently from curriculum projects which focus on process goals. One example is the Social Studies Inference Test, a paper-and-pencil test which was developed as a criterion measure for the study, Thinking in Elementary Children (Taba, Levine, and Elzey, 1964). Scores are obtained for the following aspects of the process of drawing inferences from data: (1) the ability to discriminate; (2) the ability to generalize; (3) the ability to recognize the limits of data and to refrain from overgeneralizing; and (4) the tendency to make errors which contradict what the data suggest. Taba et al. report a significant relationship between childrens' performances in classroom discussions and scores on the Social Studies Inference Test. This test was well designed for a study which focused on patterns of group interaction in an intact classroom. However, no systematic attempts were made to determine how accurately the test could characterize the individual child on the four processes. A test must address itself to this issue if it is to be used to assess the impact of a program on individual children.

Another curriculum project which focuses on process goals is Science - A Process Approach, designed by the American Association for the Advancement of Science Commission on Science Education (1969) for elementary science. The Process Instrument was built to assess longitudinal development of thirteen process goals such as observing, communicating, inferring, formulating hypotheses, and experimenting. The basis for this test is the sequence of behavioral hierarchies which is the core of this program. This instrument is administered individually. The questions are available in booklet form, and materials are specified which are required for certain questions. Highly standardized procedures are developed for administering and scoring the test. Although the Process Instrument seems to be well related to the goals of the curriculum, a test of this type does not seem appropriate when concepts must be integrated and applied to a new situation.

A final example of a test of the process goals of a curriculum is Questest, an individually-administered test for the Inquiry Training Program (Suchman, 1962). In Questest, a student is presented with a problem episode on film; he is then required to explain the phenomena depicted. He gathers data by questioning the examiner, and his verbal protocol is analyzed to evaluate several processes; questions asked, fluency of questioning, frequency of various categories of questions, and overall plan or strategy of questioning. Because of the close correspondence between Questest and the criterion, it appears to be a

highly valid test. However, one limitation of the protocol analysis is its failure to consider groupings or patterns of questions asked. A further drawback is the lack of highly standardized testing procedures; the examiner responds to the specific set of questions generated by the child. The rules for the examiner's behavior have not been made explicit, nor has the interaction between child and examiner been well controlled.

This brief review of some instruments used to assess process goals clearly indicates that techniques are available to researchers who want to study the dynamics of complex behaviors rather than their end products. Although it is also evident that many methodological questions remain to be answered about these techniques, the tests discussed are important first steps for dealing with this very difficult measurement problem.

The final component of the model is the verbal protocol or "thinking aloud" technique. Because different mental processes can produce the same answer, this technique has been proposed for making such processes overt and subject to our scrutiny. Typically, the child makes a verbal report while he engages in the criterion behavior, and the resulting verbal protocol is analyzed as the dependent measure. The following is a representative sample of studies which have successfully employed this methodology.

An early use was made by Durkin (1937) in comparing trial-and-error versus insightful approaches to problem solving. He required his subjects to verbalize their plan of attack before permitting them to manipulate the pieces of a puzzle. In this way he separated those who were unaware of the goal from those who were verifying clear-cut hypotheses about the relationship of the pieces to the goal.

Verbal protocols have been frequently used to analyze the sophisticated problem solving of adults. For instance, Bloom and Broder (1950) presented college students with a variety of problems which could be solved if they were systematically analyzed. From the analysis of these verbal reports, the authors developed a check list of characteristics of good and poor problem solvers. Duncker (1945) and Paige and Simon (1966) have also used verbal protocols to study the solving of complex mathematics problems.

Although the vocabulary of children is limited compared to that of adults, verbal protocols have also been used to study their thought processes. For example, much of the work of Piaget (1954) on the development of thought is based on verbal protocols of children. Riber, Murphy, Woodcock, and Black (1952) have used this technique to study problem solving in 7 to 8 year-old children. They presented tasks to individual children in three stages: (1) the task was shown to the children, who were asked to say what materials were needed to solve it; (2) they were given the necessary materials and were asked to tell how they planned to go about the task; (3) they were then allowed to complete the task. Their protocols were analyzed to determine how much they depended on manipulation of the materials to solve problems.

Inquiry in 10 to 14 year-old children has been examined by Donaldson (1963). She had children "think aloud" to determine their method of attacking problems and to detect their errors. She used these results to devise a theory to predict the discriminating power of new, untried test items.

Although, as Donaldson cautions, we must be aware that verbal reports cannot tell us everything, the above studies suggest that they do provide a great deal more relevant information than other methods, such as the tab format. Of course, when the protocols have been collected, the information of interest to the researcher is not in a usable form. The detailed, painstaking analysis of the reports is an integral part of the instrument, and no well-defined rules have been developed for such analyses.

This analysis of these three methodologies suggests that they can be blended into a model for assessing complex educational goals. The result is a performance test which is closely related to the criterion behaviors of educational programs; the students' verbal reports permit a comprehensive look at the process goals of such programs. Although it appears that all three of these methods have never before been combined into a single instrument, such a model seems appropriate for the purposes which any test of complex educational outcomes must satisfy.

TESTING THE MODEL

To determine the usefulness of this model, I used it to develop a diagnostic test to assess each student's performance in an individualized inquiry lab in mathematics. The lab was developed at the Learning Research and Development Center at the University of Pittsburgh, and a preliminary version was field tested last year. In the lab, one or more fifth-grade students chose from an array of projects and devised and carried out an approach to that project. A project included written or taped directions, required materials and equipment, and resource materials such as books and films.

In consultation with the curriculum development team for the mathematics inquiry lab, five process goals were chosen which seemed to represent best the complex goal of inquiry. These processes were: comprehension of what is required to work on a project, analyzing the project and planning a strategy for approaching it, execution of one or more plans, self-evaluation of performance on a project, and reporting results on a project. Thus the instrument assessed the dynamics of mathematics inquiry, not just its end products.

The universe of criterion behaviors for each process goal was carefully defined by a logical task analysis. An example of the universe for the process goal "analyzes and plans the task" can be found in Appendix A.

Three tasks were used to sample the criterion behaviors, because it has been shown that a number of smaller tasks in a performance test is more reliable than a single task (Adkins, 1951). Five criteria guided the development of these tasks:

1. To detect a wider range of inquiry competencies corresponding to different levels of inquiry sophistication, each task could be solved in more than one way.
2. The difficulty level varied across the three tasks.
3. Because children often depend heavily on the manipulation of concrete materials when solving problems (Biber et al., 1952), and because the amount of manipulation a task permits appears to be related to its intrinsic interest for children (Sears, 1966), each task contained materials and apparatus to be manipulated.
4. Each task had only one correct solution to provide "a structured and closed frame of reference...(which) facilitates the difficult task of observing and analyzing [p. 67]." (Donaldson, 1963)
5. To increase the generality of the test beyond the three tasks used, each was capable of structural analysis. The structural analysis used is discussed by Polya (1957) as the appropriate analysis for "problems to find (whose aim) is to find a certain object, the unknown of the problem [p. 154]." Each problem was separated into three principal parts: the unknown, the data, and the conditions. The complete procedure for this analysis is discussed in Appendix C with the guidelines for using the scoring key. A description and structural analysis of Car Trip, one of the tasks presented to the child, is found in Appendix B.

The "thinking aloud" technique was used to make the five process goals overt; the analysis of a tape of each child's verbal protocol served as the dependent measure.[1] In order to break down the five process goals for scoring, a total of 49 behavior categories were logically developed; these categories defined the major components within each process. The child received a score for his responses on each of the three tasks as those responses related to the categories. Because there were unequal numbers of categories within process goals and because some categories applied more than once, a percentage score was used. A copy of the guidelines for using the scoring key is found in Appendix C
A study of the reliability of the scoring indicated 97% agreement between independent scorers. The five process goals will be briefly described below.

A: Comprehension of the Task. This score measures the ability to understand independently the unknown to be found in the task and to keep it in mind throughout the task performance. Until the unknown is identified and accepted as legitimate by the child, fruitful inquiry into the task is very unlikely. Two assumptions must be made about the tasks presented to the child: (1) the information about the task is clearly presented in language which a fifth-grade child can understand; (2) the problems are interesting and meaningful.

[1] A sample protocol for the Car Trip problem is found in Appendix D.

B: Analysis and Planning for the Task. This process goal requires the child to recall and apply formerly-acquired knowledge to a new situation or to rearrange data to discover patterns in them. The categories for this goal are designed to reflect the processes of analyzing and planning rather than the correctness of the plan conceived. Therefore, the score for this goal measures the child's ability first to identify the relevant data and conditions of the task and then independently to incorporate all of them into one or more procedures which relate them to the unknown. A child achieves a higher score if a plan is stated first rather than simply emerging from a period of aimless trials. The completeness of the plan and procedures to be performed also contributes to a higher score.

C: Execution of the Plan. The score on this goal measures how well the child correctly carries out the operations outlined in his plan. Frequently these involve accurately collecting data, computing, and measuring. Although it seems that some method would be desirable for weighting this score according to the quality of plan executed, no such procedure was used in this study.

D: Self Evaluation of Performance. The score on this process goal indicates how well the child checks his plan of attack, his procedures, and his answer against the demands of the task; it indicates whether he redirects his efforts when he is not progressing toward the goal of the task. It also assesses whether he verifies his answer by some other method and whether he provides some reasonable argument to support his answer.

E: Reporting Results on the Task. This score measures the ability to report completely the steps followed in finding the unknown, as well as the unknown itself. It asks the child to review the methods he has used to reach his solution and to see how the parts of the problem are related. By providing for this kind of consolidation, this procedure probably improves the transfer of methods to other problems of the same type.

The three tasks were administered according to standardized procedures to a representative sample of ten children who were participating in the mathematics inquiry lab. This try-out of the Performance Test of Mathematics Inquiry was intended to answer three major questions: (1) Does the test provide valid, reliable information about the inquiry characteristics of these children? (2) Is this kind of instrument feasible and appropriate for assessing complex educational goals? (3) Do the results from the test suggest changes in the curriculum, which was still under development? As I expected, the testing of this model did produce many insights into these three areas of concern. Let me now share some of these with you.

One of my first problems was to find a method for assessing the reliability of the test scores. Because there were only three tasks, I could not compute the split-half reliability from the length of the

instrument. Instead, I decided to view the three tasks as parallel tests in the sense that they were designed to sample the same behavioral outcomes. The relationship between the scores obtained on the three tasks was then used to provide an estimate of the reliability. Because inspection of the percentage scores for the group suggested that they would not meet the assumption of homogeneity of variance, a non-parametric statistic, the Friedman two-way analysis of variance (Siegel, 1956), was used. This test determines if the ranks of the students on the three tasks were similar. The results indicated that the ranks were similar on all the process goals except "Analysis and Planning for the Task". An inspection of students' scores on the three tasks for that process goal revealed that their ranks on the Car Trip problem were different from their ranks on the other two tasks. The need for better techniques for assessing the reliability of this kind of test is obvious. Although the results from the Friedman two-way analysis of variance do not indicate adequate reliability for making fine discriminations among individual students, inspection of the scores suggests that the test does classify students more grossly on the various process goals and pinpoints areas in which they are particularly strong or weak. Such information is certainly useful for making some kinds of instructional decisions. If more precise information is desired, the reliability of the test could be improved by using more tasks which encompass a much wider range of difficulty. This would, of course, increase the cost of the test in terms of time and personnel and would require better methods for determining difficulty level of tasks.

My second consideration was the validity of the test. Since the child is required to perform tasks which are very similar to the criterion, the test appears to have high sampling validity; also, the scoring key for the protocols describes the kinds of inquiry-related behaviors we expected to see in the math inquiry lab. In addition, I used several different methods to collect information on the empirical validity of this instrument.

The first method I used was observations in the classroom. Each child who was tested was observed for several five-minute time samples; the order of the samples was randomized so that each child would be observed at different times during the class period. Over four months a total of 44 observations was made. Because of the difficulties involved in developing a structured schedule when the relevant inquiry behaviors in the classroom had not been clearly identified, I used unstructured observations. A random sample was chosen of matched pairs of observation records produced by independent observers; an index of agreement was calculated and the pairs of records were found to agree 89% of the time. Prior to the observations for this study, observations were made and categories of inquiry-related behaviors were developed inductively for each of the five process goals. This coding scheme attempted to delimit positive and negative inquiry behaviors in the classroom and to relate them to the five process goals to be assessed by the performance test (plus a sixth goal, "Involvement with the Task"). Appendix E contains a sample of positive and negative

behaviors for the process goal, "Analysis and Planning for the Task". This coding system was then used to code a new set of observations, but adequate agreement could not be reached between coders who were using it.

An alternative analysis of the observations was then made; a positive, negative, or "cannot say" rating was assigned to each of the 44 observations for each of the six process goals. The frequencies of the ratings were tallied for each child and transformed into a percentage for comparison with the results of the Performance Test of Mathematics Inquiry. The percentage of agreement between raters of the observations ranged from 70% to 87%. A high percentage of "cannot say" judgments and the limited number of observations made in the classroom raised serious questions about the use of these observations to establish the empirical validity of the test. However, the observations did serve two important purposes. First, they provided us with information on positive and negative inquiry-related behaviors; this represents one of the first attempts to state behaviorally the kinds of things a child does when he inquires in an instructional setting such as the math inquiry lab. Second, they point to characteristics of inquiry (and, very probably, of other complex behaviors) which should be considered in future observations. In particular, I found that there were too many behaviors to observe precisely; further observations should definitely be more focused.

In addition to the observations, empirical data were collected for each child on three additional measures. The first measure was a product score which indicated the number of tasks on which the child reached a correct solution. A slightly positive relationship was found between scores on process goals and the product score; this was expected, since the ability to reach a correct solution depends in part on the processes used to reach that solution. However, I did find some instances in which a correct solution was found using an incorrect plan and procedure. This is further evidence for the need for assessing the components of complex behaviors to determine how the child proceeds.

A second additional measure was a classification of the plans used by the child in working on each of the three tasks. The strategies used across the entire group of children were ranked according to level of sophistication. Appendix F contains a classification of the strategies used in the Car Trip problem. The three levels included: (a) a child correctly relates all the relevant factors, applies a rule, or seeks a pattern; (b) a child incorrectly relates all the relevant factors, applies a rule, or seeks a pattern; (c) a child fails to relate all the factors, introduces irrelevant factors, violates a condition of the task, or does not seek patterns. I found that over half of the plans used to solve the three tasks were at the lowest level. This result seems to be reflected on the Performance Test of Mathematics Inquiry in the poor scores of the children on the process goals involving planning and self-evaluation.

The third validity measure was an analysis of the nature of the

children's errors according to the model developed by Donaldson (1963). This can be interpreted as a measure of the seriousness of the child's difficulties in the inquiry situation. The first and most serious type of error is the arbitrary error, which occurs across tasks and which arises from a failure to attend to the givens of the task. The second type, the structural error, is specific to a given task in that the child fails to appreciate the relationships involved in the problem. The least serious is the executive error, in which the child fails to carry out the required manipulations. These kinds of errors are reflected in the behavior categories for the various process goals. In general, high scores on the Performance Test of Mathematics Inquiry were associated with executive errors, and low scores were associated with either structural or arbitrary errors.

Additional validation studies of this instrument are needed if it is to be used with a high degree of confidence. However, the evidence from the measures used here suggests that the test is indeed tapping some of the important behaviors involved in the process of mathematics inquiry.

FEASIBILITY AND IMPLICATIONS

Let us assume that in principle tests developed from the model I am proposing can be adequately refined to meet rigorous scientific criteria. I would now like to address myself to other considerations: are such tests feasible for use in schools, and what kinds of information can they provide?

Compared to paper-and-pencil tests designed for group administration, the use of a test such as I have just described is quite expensive; this is true because it must be administered individually, and because the analysis of the dependent measure is quite complicated. There are at least two possible reasons why a test developed from this model is so expensive. First, as yet we have little empirical data on the nature of many complex learning goals and few well-developed methods for determining the validity and reliability of such an instrument. Second, perhaps more complicated (and hence more expensive) methodologies are necessary if we are to capture the complexities of these kinds of outcomes.

One means of cost reduction, which hopefully will soon be available, is computerized analysis of verbal protocols. Another way to reduce costs is to consider carefully the kinds of information desired from the test; accurate, precise data on the performance of every child is much more expensive than more global data about group trends. A third possibility is to collect information on only the critical process goals. For example, my study suggests that planning and self evaluation are the two most important processes of mathematics inquiry; performance in these areas seems to discriminate well between the more and less capable inquirers. If further research supports this, we may

need to collect only that information.

Besides the cost factor, there is another limiting aspect of this model: it demands that the child be able and willing to share his "thoughts" with us. In many school situations these conditions may not hold. One thing I did find is that some children in my study who lacked verbal facility depended on the materials and apparatus to demonstrate what they were "thinking".

Although the model does have these limitations, it can serve many useful functions. For example, some of the information it provides does not seem to be reflected in paper-and-pencil tests of ability and achievement or in school performance records. Such information may very well suggest the kinds of instructional intervention which are needed. I found that the results from the Performance Test of Mathematics Inquiry were not highly related to the other measures of school performance. For instance, one child who was tested consistently came up with highly efficient strategies based on careful analysis of the task, yet his performance in the mathematics curriculum and on other tests of math ability was not exceptional. However, further inspection of the data from the Performance Test revealed that although this child had no difficulty in saying what should be included in his lab report, he spent more time than any other child in writing that report. Perhaps this child has a slight motor dysfunction which has not been diagnosed but which is affecting his performance in the classroom and on paper-and-pencil tests. Obviously, steps should be taken either to alleviate his motor difficulties or to modify methods of instruction to compensate for those difficulties.

Earlier I proposed that a major advantage of this model is that it focuses attention on the processes the child uses in complex learning situations. This kind of data is certainly useful for making decisions about instruction. However, tests built from this model have additional payoff for the development of theories about complex learning behaviors; as soon as we are clear about what children do when they engage in such behaviors and what kinds of mistakes they make, we can begin to probe to find out why.

Another advantage of this model is that it seems to promote serendipity. During the development of the Performance Test of Mathematics Inquiry, I observed many unexpected events which suggested additional areas for research. For instance, I found that some of the children who received fairly low scores on the Performance Test had introduced a great deal of fantasy material into their verbal protocols; however, all of them were functioning adequately in the classroom. It would be interesting to find out if there is a relationship between the ability to inquire and the presence of such fantasy material. Another unexplained finding was that the complexity and relevance of the data to the tasks I used did not seem to affect the difficulty of those tasks.

A final area of concern that I would like to discuss with you

today is the many ways in which the use of this model can provide curriculum developers with information about the effectiveness of developing curricula and can suggest changes in those curricula.

The curriculum development team examined the group means on the five process goals of the Performance Test of Mathematics Inquiry; the scores were fairly low on B (analysis and planning for the task) and very low on D (self-evaluation). These results were generally supported by the classroom observations. The team learned that many children did not have a wide variety of strategies available in their repertoire for attacking problems. It had been assumed that the children would increase their repertoire of strategies as an implicit by-product of working on various projects under the supervision of the classroom teacher; the test results suggested, however, that more systematic instruction in planning was needed in the math inquiry lab.

The very low scores on self-evaluation revealed, among other things, that the children were unable to estimate answers. The curriculum developers had assumed that the children had learned various methods of approximation in the regular mathematics program. If this is the case, the children did not transfer those skills to the "real life" projects which were posed for them in the math inquiry lab. Again, systematic instruction about methods of estimation should be provided in the lab.

Inspection of the protocols revealed that many children seemed to believe that their ideas were not worthwhile. This was true across all levels of performance. Thus I found that if a child were asked why he did some manipulation or calculation, he immediately assumed that he had made an error which needed to be corrected. This is not totally unexpected, since many teacher-pupil contacts in the classroom involve correction of errors. However, such heavy dependence on outside authorities for feedback about the correctness of one's responses is antithetical to the spirit of inquiry. No easy solutions can be offered to this problem, but it is important that curriculum developers in the math inquiry lab have some information on how persavive this problem is.

Some of the findings from the Performance Test of Mathematics Inquiry suggested possible changes in the regular mathematics curriculum. Many of the children had problems in relating numerals to objects; because of this, the children would add 'x' miles and 'y' seconds with impunity. Clearly, more emphasis needs to be placed on the meaning of units. Another problem which some of the children experienced on the Test was an inability to handle irrelevant data. If a number occurred in a problem, they seemed compelled to include it. Apparently this difficulty is caused in part by the fact that the problems in the regular mathematics curriculum include only relevant data. It would seem desirable to include some problems containing irrelevant data in the regular mathematics curriculum.

Additional information about the curriculum emerged from the classroom observations. In the lab each child had access to all of the

projects. From the observations, I found that many children had poor strategies for choosing their project. The curriculum development team then began to provide some information about the difficulty level of each project and the kinds of skills needed to work on that project. If a child lacked a prerequisite (such as the ability to use a balance beam), information was given on ways to learn that skill.

In making the observations, I found that it was quite difficult to observe instances of process goal B (analysis and planning for the task) and E (reporting results on the task). One possibility considered by the curriculum development team was to require each child to prepare a statement of his plan before doing a project and to prepare a complete lab report when he had finished a project. Not only would this provide the team with a means for monitoring student progress, but it is also a learning experience, because the child becomes more aware of his own processes in carrying out a task.

As I indicated earlier, most of the data from the observations must be interpreted with caution. However, the data collected on "Involvement with the Task" was fairly reliable. A child received a positive rating in this category if he were working steadily on the project during the five-minute time sample. Analysis of these ratings indicated that in the fairly open environment of the lab the children were working on their project approximately 75% of the time. One goal of the curriculum team was to provide the children with "intrinsically interesting" projects. The high degree of involvement reflected by the observations suggests that in most cases this goal was being met.

In these examples, I have tried to sketch out for you some of the ways in which a test of complex educational goals provided valuable information to one curriculum development team. The model I have proposed forced the curriculum developers to be clear about their goals; it provided a framework for gathering relevant information from a very complex situation. The model also focused observations in the classroom in such a way that the curriculum developers could detect weaknesses in the system while there were opportunities to intervene and correct them. Finally, the model, if refined, could be used to monitor pupil progress in a more stable instructional system.

APPENDIX A

PROCESS GOAL: ANALYSIS AND PLANNING FOR THE TASK

Criterion: The child identifies the data and conditions of the task and incorporates them into a plan for finding the unknown or goal. The plan outlines the data to be found, procedures for collecting those data, and some method for relating the data to the goal without violating the conditions.

CATEGORY	SCORE	BEHAVIOR
1. Identifies given data	2	The child identifies all the relevant data which are given.
	1	The child identifies some of the relevant data which are given.
	0	The child does not identify the relevant data which are given.
2. Identifies data to be found	2	The child identifies all the relevant data which need to be found.
	1	The child identifies some of the relevant data which need to be found.
	0	The child does not identify the relevant data which need to be found.
3. Identifies the conditions	2	The child identifies all the conditions of the task.
	1	The child identifies some of the conditions of the task.
	0	The child does not identify the conditions of the task.
4. Questions examiner about conditions	2	The child questions the examiner about the conditions of the task before stating and/or executing a plan.
	1	The child questions the examiner about the conditions of the task while executing a plan.
	0	The child never questions the examiner about the conditions of the task and violates a condition while executing a plan.

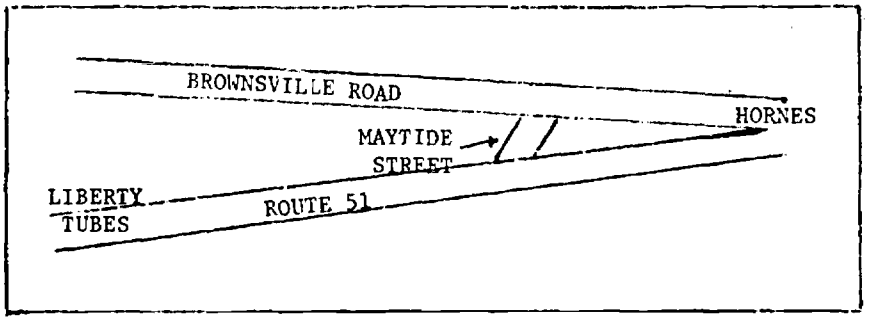
5. States plan first	2	The child states a plan before manipulating the apparatus or executing any part of a plan.
	1	The child states a plan after manipulating the apparatus or while executing a plan.
	0	The child never states a plan.
6. States plan independently	2	The child states a plan without assistance from the examiner.
	1	The child states a plan with assistance from the examiner.
	0	The child does not state a plan in spite of assistance from the examiner.
7. Plan includes relevant data	2	The child states a plan which includes only relevant or redundant data.
	1	The child states a plan which includes both relevant and irrelevant data.
	0	The child states a plan which includes no relevant data.
8. Plan includes all conditions	2	The child states a plan which includes <u>all</u> the conditions of the task.
	1	The child states a plan which includes <u>some</u> of the conditions of the plan.
	0	The child states a plan which violates a condition of the task.
9. Plan relates data and conditions to the unknown	2	The child states a plan which relates <u>all</u> the relevant data and conditions to the unknown or goal.
	1	The child states a plan which relates <u>some</u> of the relevant data and conditions to the unknown or goal.
	0	The child states a plan which does not relate the relevant data and conditions to the unknown or goal.
10. Complete plan	2	The child states a plan which includes <u>all</u> of the following: any data to be found, procedures for collecting the data,

		and some method for relating the data to the unknown.
	1	The child states a plan which includes <u>some</u> of the following: any data to be found, procedures for collecting the data, and some method for relating the data to the unknown.
	0	The child states a plan which does not include the following: any data to be found, procedures for collecting the data, and some method for relating the data to the unknown.
11. Complete procedure	2	The child states a plan which includes procedures to be followed. The child states the act to be performed, the object of the action, and the attribute of the object upon which he will act.
	1	The child states a plan which includes procedures to be followed. The child states the act to be performed and the object of the action.
	0	The child states a plan which includes procedures to be followed. The child states the act to be performed but does not include the object of the act.
12. Uses relevant information in a new plan	2	The child fails to reach a satisfactory solution and changes his plan. He incorporates <u>all</u> relevant information from his previous failure into the new plan.
	1	The child fails to reach a satisfactory solution and changes his plan. He incorporates <u>some</u> of the relevant information from his previous failure into the new plan.
	0	The child fails to reach a satisfactory solution and changes his plan. He does not incorporate any relevant information and/or incorporates irrelevant information from his previous failure into the new plan.

APPENDIX B

CAR TRIP

Mr. Smith is traveling in his car along Brownsville Road. It takes him four minutes to go the distance from Hornes to Maytide Street. This distance is marked in black lines on the track. His car moves at thirty miles per hour and does not stop. Mr. Smith then turns down Maytide Street and drives to Route 51. His car moves along Route 51 at thirty miles per hour without stopping until he reaches the Liberty Tubes. This distance is also marked in black lines on the track. The scale for this track is shown in the corner above Hornes. How could you find out how many minutes it will take Mr. Smith to travel the distance from Maytide Street to the Liberty Tubes along Route 51?



CAR TRIP APPARATUS

CAR TRIP PROBLEM -- STRUCTURAL ANALYSIS

1. The unknown:

   The time required for an object to travel from Point C to D.

2. The conditions:

   a. The object must travel in a straight line between the two points.

   b. The object must travel at a constant speed.

3. The data:

   GIVEN:

   a. The object travels between points A and B in 4 minutes (relevant).

b. The object moves at 30 m.p.h. between both Points A and B and between Points C and D (relevant).

c. A scale relates the dimensions of the apparatus [in inches] to those of the real object [in miles] (redundant).

MIGHT BE FOUND:

a. The distance between Points C and D is 4 times the distance between Points A and B (relevant).

b. The scale relationship 3" = 1 mile (redundant)

c. The distance between Points A and B is 6" (redundant).

d. The distance between Points A and B is 2 miles (redundant).

e. The distance between Points C and D is 24" (redundant).

f. The distance between Points C and D is 8 miles (redundant).

g. The object travels 6 inches in 4 minutes (redundant).

4. Most efficient strategy:

Find the time to travel one known-distance unit. Find the number of known-distance units in the unknown distance. Multiply that number by the time required for the one known-distance unit. The product is the unknown time.

APPENDIX C

SCORING KEY

PERFORMANCE TEST OF MATHEMATICS INQUIRY

I. GENERAL INSTRUCTIONS

When using this key, the scorer should follow these guidelines:

1. Become completely familiar with the key on the following pages by reading it carefully. Make sure that you completely understand the criterion behaviors for each of the five process goals.

2. Analyze each task on the performance test to determine the goal or unknown, the conditions, the data which are given or which might be found, and the most efficient rule or strategy for finding the unknown. Decide which data are relevant, redundant, and irrelevant with respect to that rule or strategy. (Section II defines these terms.)

3. Arrange your work so that you score all the protocols for one task before proceeding to the next task. By doing this, you should not be influenced by the examinee's performance on other tasks. For this same reason, the name of the examinee should not be shown on the protocol.

4. Before scoring a protocol, read it completely so that you are familiar with the child's approach to the entire task.

5. Determine the beginning and the end of each plan and mark all such points on the protocol. Then indicate whether or not each plan has been executed.

6. Score the protocol for each process goal in order from A through E. Each category should be assigned a two, one, zero, or "does not apply".

7. On those process goals where more than one plan must be scored, completely score the first plan before starting to score the next one.

8. When you begin to use the key, score one protocol for each task by yourself and then have these same protocols scored by a second person who is familiar with the key. Compare your scores for each category within each process goal. Discuss any disagreements and repeat this process on a new set of protocols until you both agree on the scoring.

9. Additional rules and further clarification of specific categories are provided in Section III. The scorer should also be

familiar with these.

II. DEFINITION OF TERMS

The following alphabetical list provides a more precise definition for several terms used in this scoring key. These definitions should be used consistently in order to avoid confusion and to increase the reliability of the scoring.

Checks plan
: The child examines the various parts of his plan and matches one or more parts of it with some part(s) of the problem situation.

Conditions
: Those terms of a problem which set the limits within which the solution to the problem must lie and which stipulate any prerequisites that must be satisfied before the problem can be completed. Conditions may disallow certain plans of attack.

Executes a plan
: The child carries through a plan which he has stated or engages in some behavior which appears to be related to solving the problem.

Failure
: The child decides that a plan or the solution reached by a plan is inadequate and abandons it for some other plan. His decision may or may not be right.

Goal
: See "unknown".

Identifies
: Mentions explicitly in the statement of a plan or uses in a plan without mentioning explicitly.

Irrelevant data
: Facts and objective information which cannot be used to reach the correct solution to a problem.

Most efficient rule or strategy
: A rule or strategy that requires that the child perform the fewest number of behaviors and that uses the least amount of data.

Plan
: Any method, procedure, or object for finding the unknown of the problem which is stated by the child or implicitly followed

by him It need not be logical, detailed, well developed, skillful, or determined beforehand.

Protocol

A written record of the child's performance on a task. It should contain three parts: (1) all verbal communications between the examiner and the child; (2) all task-related behaviors of the child which have been noted by the examiner; (3) all written materials produced by the child.

Redundant data

Facts and objective information which can be used to reach the correct solution to a problem but which are unnecessary if the most efficient rule or strategy is used.

Relevant data

Facts and objective information which must be used to reach the correct solution to a problem by means of the most efficient rule or strategy.

Unknown

The part of a problem which must be found. Finding it is the aim or purpose of the task.

APPENDIX D

SAMPLE PROTOCOL

The examiner reads the Car Trip problem to the child.

Examiner: Do you have any questions?

Child: Okay, he goes from Hornes to Maytide Street in four minutes. His car is going 30 miles an hour, it doesn't stop...okay... So that's 4.

Examiner: First, I want you to tell me what the problem is.

Child: You want me to find out how long it will take him to get from Maytide Street to the Liberty Tubes.

Examiner: Before you start working on it, can you tell me what you think you are going to do?

Child: Well, I'll take 4 minutes into 30 miles per hour. That will tell me how many minutes it takes him to go from... He goes from Hornes to Maytide Street in 4 minutes, traveling at 30 miles per hour and his car doesn't stop. So I'll just take the 4 minutes and measure how far from Hornes to Maytide Street is. I'm going to take 4 minutes into 30 miles an hour. I already did that... [The child takes a sheet of paper and duplicates the scale on the apparatus and uses this to measure] It says here he goes to Maytide Street in 4 minutes. His car is going 30 miles an hour and he does not stop. Mr. Smith drives down Maytide Street to Route 51. The car moves along Route 51 at 30 miles per hour without stopping. So he went 2 miles in four minutes, 30 miles an hour...So...Hornes to Maytide Street...That's right... So...he went 2 miles here. I'll keep measuring from here (Maytide Street), every two miles it will be 4 minutes since he is going 30 miles an hour and he didn't stop.

Examiner: Do you want to do that?

Child: [The child measures the distance from Maytide to the Liberty Tubes with the movable scale. The child's finger is used to mark off these segments.] 2 miles. 3 miles. 4 miles. 5 miles. I'll just do two times that... This would be two miles. Right? Two.

Examiner: How did you get two here?

Child: Well, this was two miles here [the scale]. I don't know what I'm doing. 4 miles. 2 miles. I'm getting mixed up. I thought this was a two miles scale. I was thinking about this. He traveled 2 miles from Hornes to Maytide. He's supposed to be traveling along here. The scale is only one mile. I have to start over. 5 miles, 6 miles, 7 miles, 8 miles...8 miles, 2 into 8 goes about 4 times. 2, it says Mr. Smith is traveling in his car... So that would be 2 miles equals 4

minutes. So that would be 4 times 4 is, uh, 8, 12, 16. It would take him 16 minutes.

Examiner: So your answer is 16 minutes? Can you think of another way?

Child: No.

Examiner: Do you think that is the right answer?

Child: Yes.

Examiner: Would you like to write a lab report for me then?

Child: What kind of paper should I put? [for the movable scale]

Examiner: Scratch paper... Can you tell me again what the problem was?

Child: To find out how many minutes it took him to drive from Maytide Street to the Liberty Tubes.

APPENDIX E

OBSERVATIONS CODING SCHEME

PROCESS GOAL B: ANALYSIS AND PLANNING FOR THE TASK

The child plans and analyzes:

1. The child correctly states the conditions of the task.

   a. the child states he will try to meet those conditions.

   b. the child states he will not try to meet those conditions.

2. The child incorporates all the relevant conditions and data of a task into his plan and procedures.

3. The child analyzes the task by analogy. He states that the current task is similar to a task he has done before.

   a. the child states the plan of attack for the analogous task.

   b. the child does not state the plan of attack for the analogous task.

4. The child analyzes the task by hypothesizing that some condition or relationship about the task is true.

   a. the child states an implication(s) of the condition or relationship.

   b. the child cannot state an implication.

5. The child analyzes the task by drawing a schematic picture or outline.

   a. the schematic includes all relevant task requirements.

   b. the schematic does not include all relevant task requirements.

6. The child analyzes the task by observing various features of a model of the final product.

7. The child analyzes the task by manipulating the materials and apparatus.

8. The child incorporates unnecessary requirements into his plan.

   a. the child states the requirement is irrelevant but that it corrects an error or improves the appearance of a product.

   b. the child states the requirement is relevant.

9. The child collects the materials and information needed to carry out the plan outlined in the written directions for the task.

The child does not plan and analyze the task:

1. The child does not divide the task into any parts (such as data, conditions, unknown).

2. The child asks someone else for a plan.

3. The child copies a plan from someone else.

4. The child works with someone else and follows the oral directions of this other person.

5. The child proceeds to work with the materials without exploring the implications of his behavior (its feasibility, whether the procedure is related to the task requirements, whether he has the prerequisite skills to complete the procedure adequately.

6. The child cannot state a plan when asked.

7. The child immediately begins to complete an item on the directions sheet without assembling his materials and information for the task.

APPENDIX F

CLASSIFICATION OF STRATEGIES

Car Trip Problem

LEVEL A:

1. Determine there is a constant speed for both the known and the unknown distances. Measure the distance he vent in 4 minutes. Mark off the unknown distance i1 these segments and multiply the number of segments by 4 minutes.

LEVEL B:

1. On the scale measure the number of inches that equals one minute. Mark off one minute segments along the unknown distance.

2. From observing the scale, assume that 4" = 1 mile. Measure off the unknown distance in 4" sections. 5 x 4 min. = 20.

3. Measure the unknown distance in inches and divide by 4 minutes.

LEVEL C:

1. Divide 4 minutes into 30 m.p.h.

2. Divide the unknown distance in miles by 30 m.p.h.

3. Guess about 10 minutes.

4. Try it with a real car.

5. Ask the examiner.

6. Measure the unknown distance in inches. That distance is the number of minutes.

7. Measure the unknown distance in known distance segments. Add the number of segments (4) to the distance of one segment.

8. Measure the unknown distance in inches and subtract the known distance from the result.

9. By the scale.

BIBLIOGRAPHY

Adkins, D. C., "Principles underlying observational techniques of evaluation," Ed & Psychol. Meas., 1951, 11, 29-51.

American Association for the Advancement of Science Commission on Science Education, Science - A Process Approach: The Process Instrument, New York: Xerox Education Division, 1969.

Biber, B., Murphy, L. B., Woodcock, L. P., and Black, I. S., Life and ways of the seven-to-eight year old, New York: Basic Books, 1952.

Bloom, B. S. and Broder, L. J., Problem solving processes of college students: an exploratory investigation, Chicago: University of Chicago, 1950.

Butts, D. P., "The evaluation of problem solving in science," J. Res. in Science Teach., 1964, 2, 116-122.

College Entrance Examination Board, Righting the Balance, Princeton, New Jersey: ETS, 1970.

Cronbach, L. J., Essentials of Psychological Testing. (2nd ed.), New York: Harpers & Bros., 1960.

Cross, K. P. and Gaier, E. I., "Technique in problem solving as a predictor of educational achievement," J. Ed. Psychol., 1955 46, 193-206.

Donaldson, M., A study of children's thinking, London: Tavistock Publications, 1963.

Duncker, K., "On problem solving," Psychol. Mono: Gen. & Applied, 1945, 58, Whole No. 270.

Durkin, H. E., "Trial and error, gradual analysis and sudden reorganization: an experimental study of problem solving," Arch. Psychol., 1937, No. 210.

Glaser, R., "Objectives and evaluation: an individualized system," Science Ed News, June, 1967, 1-3.

Paige, J. M. and Simon, H. A., "Cognitive processes in solving algebra word problems," in Problem solving: research, method and theory, Kleinmuntz, B., (ed.), New York: Wiley, 1966. Ch. 3.

Piaget, J., The construction of reality in the child, New York: Basic Books, 1954.

Polya, G., How to Solve it, Garden City, New York: Doubleday Anchor Books, 1957.

Sears, R., "Process pleasure," in Bruner, J. (ed.), Learning about learning: a conference report, OE-12019, Cooperative Research Monograph #15, U. S. Government Printing Office, 1966, 44-46.

Smith, E. R. and Tyler, R. W., Appraising and Recording Student Progress, New York: Harper, 1942.

Siegel, Sidney, Nonparametric Statistics for the Behavioral Sciences, New York: McGraw-Hill, 1956.

Suchman, J. R., The elementary school training program in scientific inquiry, Urbana Ill.: Univ. of Ill., 1962.

Taba, H., Levine, S., Elzey, F. F., Thinking in Elementary School Children, San Fransisco, San Fransisco State College, Cooperative Research Project, #1574, 1964.

Wechsler, D., Wechsler Intelligence Scale for Children, New York: Psychological Corporation, 1949.