DOCUMENT RESUME

ED 065 557 TM 001 702

AUTHOR Light, Judy A.
TITLE Formative Evaluation Procedures for the In-Context Development of Instructional Materials.
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE 72
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 3-7, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement; *Classroom Environment; Curriculum Development; Curriculum Evaluation; Educational Improvement; *Evaluation Techniques; Feasibility Studies; *Formative Evaluation; *Instructional Materials; Performance Factors; Student Behavior; *Student Testing; Teacher Behavior; Test Construction; Test Results

ABSTRACT

Specified procedures for evaluating materials during their in-context tryout are presented. These procedures deal with all possible causes of system failures that have been identified in the in-classroom tryout of new materials. Therefore, methods for identifying, controlling, and monitoring all factors that affect academic behavior in the classroom are described. This includes (1) the definition of classroom management rules and ways to monitor their effectiveness, (2) the collection of objective and subjective data to discover weaknesses in both the materials and the classroom environment, and (3) the ways to systematically evaluate the effectiveness of all changes made to the environment. Testing procedures were also defined. They included where to take a test, how to take a test, and how to score the test. Test performance was used to locate inadequate materials and to evaluate revised materials. Systematic formative evaluation during the in-context tryout of materials was shown to be feasible. (Author/DB)

FORMATIVE EVALUATION PROCEDURES FOR THE IN-CONTEXT DEVELOPMENT OF INSTRUCTIONAL MATERIALS

Judy A. Light

Learning Research and Development Center

University of Pittsburgh

1972

A paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 3-7, 1972

The research herein was supported by the Learning Research and Development Center supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred.

Formative Evaluation Procedures for the In-Context Development of Instructional Materials

Judy A. Light
University of Pittsburgh

Evaluation, which may be defined as the collection and use of information to make decisions about a program (Stufflebeam, 1970) can serve many roles in the design, implementation, and final assessment of an educational product. Cronbach (1963) and Hastings (1966) have discussed two major roles evaluation can serve: evaluation for decisions concerning adoption of a final product and evaluation for revisions of a developmental product. Scriven (1967) has labeled these two types of evaluation as summative and formative evaluation. Most work in evaluation has been done in summative evaluation, where the completed program is assessed, as opposed to formative evaluation, where the refinements are made before the completion of the program. The importance of the careful evaluation of materials during their construction has been widely recognized but few have attempted to outline specific guidelines for using formative evaluation in curriculum development.

There are considered to be three stages in the development of instructional materials: the pre-tryout, the in-context tryout, and field testing. The aim of the first stage, the pre-tryout, is to detect gross deficiencies in new materials. During this stage, the materials are used by an individual with

an observer in a tightly controlled environment where the observer can interact with the student. Although this stage is useful in detecting some problems within the materials, it offers little information as to the success of the materials in an operating classroom, where numerous variables can effect academic performance. This is the purpose of the second stage, the in-context tryout. Procedures for conducting the tryout during this stage are hard to find. The final stage, field testing, consists of placing the materials into many classes where few controls are possible. During this stage only summative evaluation is practical. Information comparing programs or noting insufficiencies about the program can be compiled slowly but any information as to the cause of inadequacies of specific materials is either impossible or impractical.

In carrying out the in-classroom tryout of new lesson materials, the developer must necessarily be concerned with all aspects of classroom operation that can affect pupil performance on the lessons. He must be concerned with how the lessons are used by teacher and pupil, with the degree to which specified procedures are followed, with pupil motivation, with the validity of testing procedures that are used, and with a number of other components of classroom operation. When developing a lesson, the writer assumes it is used under certain conditions. The lesson can be given a meaningful tryout only if these conditions exist. In-classroom evaluation, then, must include the study of the extent to which the necessary conditions are present.

When the person carrying out this type of evaluation obtains information such as that pupils are not mastering a given skill that is covered in a part of the lesson, he cannot immediately assume that this part of the lesson is at fault. He must be concerned with other hypotheses that could explain this lack of mastery. Is the lesson being used properly? Is the criterion test a valid one? Was the pupil motivated to learn this skill? These and other alternative hypotheses must be investigated before a decision is made to revise the lesson.

It becomes apparent that a curriculum writer should no longer be concerned only with the lesson materials. He should also define the total environment in which the materials are to be used. He should define the behaviors of the teacher and the student, the work skills each pupil is to use, the information the teacher is to give each pupil and how the necessary skills can be taught.

It is the purpose of this paper to present clearly specified procedures for evaluating materials during their in-context tryout. Because the curriculum developer should be concerned with the total environment, these procedures deal systematically with all possible causes of system failures that have been identified in the in-classroom tryout of new materials. Therefore, methods for identifying, controlling, and monitoring all factors which effect academic behavior in a classroom will be described. This will include (1) the definition of classroom management rules and ways to monitor their effectiveness, (2) the collection of objective and subjective data to discover weaknesses

in both the materials and the classroom environment, and (3) the ways to systematically evaluate the effectiveness of all changes made in the environment.

One of the major problems in doing research in curriculum development has been in selecting an appropriate design. True experimental designs have little applicability or feasibility in formative evaluation studies at this time. One of the requirements for a true experimental design is the random selection of groups (Stanley and Campbell, 1963) which can be considered equivalent on all crucial dimensions except for the treatment they receive with respect to the experimental variables. Although in formative evaluation studies random selection of students and teachers is possible, it would be difficult to maintain equivalent classroom conditions between groups. Since classroom conditions, such as teacher behavior and motivation, which effect academic performance, are heavily dependent upon the individual teacher's style, it appears impractical to assume equivalent treatments could be maintained. Also since formative evaluation is concerned with answering questions about the quality of each component of the program, an experimental design comparing two groups would offer little information about the causes of failure which are necessary for the development of a program.

Since the use of true experimental designs appears at this time inappropriate for use with formative evaluation procedures, curriculum developers can seek other types of designs

to establish causal relationships similar to the "persuasive causal interpretations made possible by experiments involving randomization" (Campbell, 1963).

Campbell and Stanley (1963) recognized that in certain natural settings the full control of the experimental variables cannot always be obtained. They have collected a group of "quasi-experimental designs" which can be used in situations where experimental designs are not practical. These quasi-experimental designs can establish causal relationships under two conditions: that the interpretations made from the collected data must seem plausible, and other plausible rival hypotheses can be eliminated (Campbell, 1963). Campbell and Stanley (1963) have listed twelve threats to validity which form a list of probable rival hypotheses. Certain quasi-experimental designs control for some of these sources of invalidity. In other designs they form probable rival hypotheses which have to be considered as possible alternative interpretations.

Sidman (1960) suggests there are only two criteria, reliability and generality, which should be considered in accepting or rejecting data. Reliability can be established by repeating similar experiments to determine if they yield the same results. Several ways of establishing reliability through replication are suggested: inter-subject direct replication, intra-subject direct replication, and systematic replication. Generality can be established by finding similar results under different conditions. Sidman advocates the use of systematic

replication where the experiment is repeated under different conditions instead of direct replication which requires all subjects to be treated alike except for the independent variable in question (Sidman, 1960, p. 111). If similar results using systematic replication are obtained, evidence supporting both reliability and generality are attained. If systematic replication fails, then the original experiment must be directly replicated; therefore, systematic replication is only sensible when there is a great deal of confidence in the techniques to support using the data as a basis for performing new experiments.

Both Campbell and Stanley and Sidman are suggesting similar strategies for using non-experimental designs. Campbell and Stanley suggest the use of the rejection of alternate hypotheses to establish causal relationships. The designer must be cautious in controlling sources of invalidity. Elminating these threats to validity increases the strength of the design by eliminating rival hypotheses. Sidman suggests using evidence of generality and reliability resulting from systematic replication to establish causal relationships. The more similar the results found in different settings, the more confidence is gained in establishing causal relationships between variables. Campbell and Stanley (1963) also suggest the need for systematic replication: "The experiments we do today, if successful, will need replication and cross-validation at other times under other conditions before they can become an established part of science, before they can be theoretically interpreted as a part of science."

The designs proposed by these authorities can be thought of as having much to offer the formative evaluator. Design specialists caution "because full experimental control is lacking, it becomes imperative that the researcher be thoroughly aware of which specific variables his particular design fails to control" (Campbell and Stanley, 1963). Instructional materials specialists (Lumsdaine, 1964; Merkle, 1964) stress that any conditions which can affect a program be specified, and either controlled or eliminated. Therefore, formative evaluation can be successful if all factors which can affect the program are specified and causal relationships can be established through the elimination of rival hypotheses.

During the formative evaluation of a developing program, the curriculum designer and evaluator seek ways to improve the instructional system. Their task of seeking direct cause-and-effect relationships between instruction and student success can be a slow and often unproductive activity. One of the needs in formatively evaluating an instructional system is a method for making rapid decisions and improvements in the instruction.

It appears useful to apply Platt's (1964) method of strong inference to the area of formative evaluation of instructional systems. Strong inference is based on the exclusion of alternate hypotheses or explanations. Alternatives which cannot be excluded are considered to establish causal relationships only until they are disproven. At any time another explanation can be found which cannot be disproven. In the area of curriculum

development a major problem is acquiring sufficient evidence to prove why materials do or do not work. The number of variables and their interactive effects present in a classroom makes it difficult to tightly control the situation. The application of strong inference as a method for formative evaluation would help in overcoming certain problems created by working in an on-going classroom and providing procedures for making rapid improvements. The evaluator could concentrate on the inadequacies of the materials by asking "why did these materials not work" and listing as alternate hypotheses all variables which potentially could contribute to failure. "Experiments" could then be designed and carried out individually to exclude each alternative. If all listed alternatives are rejected, the evaluator's task then becomes one of seeking other alternative explanations. If one alternative explanation cannot be eliminated, it is momentarily accepted as the "cause" of failure. The instructional system is then modified according to the accepted hypotheses until another failure results, starting the cycle over again.

Another major problem is designing procedures which allow the evaluator to be simultaneously concerned with all identifiable aspects of a classroom environment which can effect student performance. If an evaluator is only concerned with some aspects of the classroom, it would be difficult to establish cause-and-effect relationships between variables since the crucial variables may not be among those he has selected to be concerned with. Once an evaluator considers all aspects of a

classroom as possible causes of failure, he can systematically disprove their effects.

From previous classroom experience the following categories of variables were found to influence academic performance: teacher behavior, pupil behavior, testing behavior, and classroom management procedures. Two methods to control the influence of these behaviors were found effective. The variables were either controlled by defining and enforcing specific rules resulting in a systematic influence on student performance or by monitoring the effects of certain variables and then considering them as possible causes of failure.

Since the major concern of this work was with improving instructional materials, teacher behavior, testing behavior, and classroom management, procedures were explicitly defined in order to eliminate their effects on the student's academic performance. The teacher was required to abstain from any tutoring during the class. This restriction was placed on his behavior to insure that what was learned was learned from the instructional materials being evaluated rather than from the teacher's tutoring. Students were instructed to redo the teaching pages in the instructional materials if they had any difficulty with the materials. The students eventually learned that the teacher would not tutor them and, therefore, stopped asking for help.

Since student motivation can effect academic performance, the teacher's task during class was to attempt to maintain high motivation among the students. Both teachers used in

this developmental work had previous experience in using behavior modification techniques to motivate students. Their behavior during class was that of a reinforcer. Praise was consistently given for passing tests, working hard, using good study skills and following rules.

Classroom management rules encompassing all aspects of the classroom were defined and given to the students. Adherence to these rules were strictly enforced by the teacher. These rules were created and enforced to insure that systematic procedures would be followed by the teachers and students. They included what to do if a student failed a test, what to do if a student was having difficulty, how to find the appropriate materials, and what test to take when.

Testing procedures were also defined. They included where to take a test, how to take a test, and how to score the test. The main purposes of these rules was to insure that the student's work was his own. No help from the teacher, other students or instructional materials was allowed once a student entered the testing area. If any rule concerning proper testing procedures was broken, the test was voided and the student was required to take an equivalent test.

In order to insure these procedures were followed, an observer was always present in the classroom. Any discrepancy, such as a pupil received help during a test, was noted. This informal data was used in hypothesizing test failures. Also, if situations were noticed by either the teacher, observer, or

evaluator which could improve the control of some variable, classroom rules were altered.

One major variable, student behavior, was not controlled by the evaluator or teacher. Because one goal of the program was to increase self-evaluation skills, appropriate pupil behavior was only defined for the students. Deviations from the appropriate procedures were always considered as a possible cause of failure. Pupils were instructed to do one page at a time, score it, and then correct their errors, follow the exact order of the pages as they were assigned and independently work on their materials. When a student failed a test, any one of these student skills could have contributed to the cause of failure. Ways of detecting the misuse of these student skills were thus included in the procedures for locating the cause of failure.

There were many benefits derived from identifying and controlling the variables which can effect academic performance. Certain non-academic causes for failure were eliminated as possible causes of failure. Once students learned the appropriate classroom rules, they stopped "trying to beat the system." They found the only way to pass a test was to learn the materials. Because teacher behavior was kept consistent in praising good work, students did not use the ploy of doing poorly to get teacher attention.

The strict rules of behavior for taking a test and the non-tutoring of students insured the evaluators that the

students' work was an accurate representation of the instructional materials worth.

Finally, the total environment, although subject to revisions, was definable. A precise description of how the materials were used was possible.

The purposes of formative evaluation are to provide information of how to improve curriculum materials, make necessary revisions in the materials, and assess their effectiveness. Before curriculum materials can be improved, a systematic process for identifying inadequacies should be identified. The curriculum used to develop these procedures contained tests for each of its objectives. Since students took these tests frequently, their test performance after each objective was a useful unit to use for locating inadequate materials. The use of test performance was also useful because the tests were used as the behavioral definition of each objective. Materials were supposed to be designed to teach students the skills necessary to pass the test. These procedures were designed to locate materials which were not adequate.

The major assumption of this model in using inadequate test performance to locate poor instructional materials is that the cause of poor test performance can be identified by systematically examining pupil performance on the instructional materials. Once a reasonable cause of failure is located, materials can be immediately revised to eliminate the identified

weakness and evaluated by assessing the same student's performance on an equivalent test.

These procedures require the evaluator to always objectively test his revisions. This use of an equivalent form of the failed test to assess the revisions provides the evaluator with immediate feedback as to his success in hypothesizing a cause of failure. If an inappropriate cause of failure is chosen to base revisions upon, the student should not pass the equivalent test which forces a repetition of the entire process.

Test performance is used within this model to locate inadequate materials and to evaluate revised materials. After each class all tests taken during the class period that day were examined. As long as a student's test performance was not considered inadequate by the evaluator, the instructional materials for that objective were not evaluated. All tests where the student did not pass every test item were considered inadequate test performance. All materials used by the student to teach the objective were gathered for intensive examination by the evaluator and teacher. For each failure, the question was asked, "Why did this student fail the test associated with these materials?" The successful use of this model is based on systematically locating and testing each hypothesized cause of failure. To accomplish this, answers to these five questions were always sought by the evaluator in order to identify a cause of test failure:

1. What was similar about the problems missed on the test?

2. How did the items missed differ from those items passed on the test?
3. Where in the instructional materials were these items presented?
4. What in the instructional materials caused the student to fail the test?
5. How can the hypothesized cause of failure be experimentally proven?

The use of these procedures to improve instructional materials appears to be most clearly explained through the use of examples. For each example a copy of the student's test is always presented. In all examples the handwritten responses are the student's answers. Those marked with an X are incorrect.

EXAMPLE I - FIGURE I

Write in the missing numbers using the associative principle.

(4x2)x5 = 4x(2x 5)
= 4x 10
= 40

2x(4x8) = (2x4)x 8
X = 32 x 8
= 256

(9x3)x6 = 9x(3x 6)
X = 27 x 18
= ....

6x(7x4) = (6x7)x 4
= 28 x 42
X = ...

The student missed three questions on the test. After class the student's test and materials were gathered for

analysis. The five questions, stated previously, were asked.

1. What was similar about the problems missed on the test?
   a. The student always made the first error on the second line of the problem.
   b. The errors appear to be systematic. The pupil always puts the product of the multiplication problems within both sets of parentheses from the first line into the blanks on the second line.
2. How did the items missed differ from those items passed on the test?
   a. The one item passed had one numeral, a 4, already written in the second line.

Because the single problem passed by the student contained an additional cue, namely the numeral four, the evaluator hypothesized that the student probably had not learned what the associative principle was from the instructional materials. The model requires the evaluator to look through the student's materials to identify why the student did not learn the appropriate skill. Our instructional materials were designed so that the last page before a test is identical in content and format to the test. Since the pages, upon inspection, appeared to explain the associative principle clearly and the student had completed the pages correctly, attention was focused on the last page before the test. Examination of the last page, appearing in Figure II, led us to a hypothesis on the cause of failure.

EXAMPLE I - FIGURE II

This is the last page before the test.

Multiplication is associative:

$$(8x2)x2 = 8x(2x2)$$

$$16\ x2 = 8x\ 4$$

$$32 = 32$$

Write in the missing numbers and solve the equation using the associative principle:

$$(3x2)x5 = 3x(2x\ 5\ )$$

$$6\ x5 = 3x\ 10$$

$$\_\_\_ = \_\_\_$$

$$(3x9)x4 = 3x(\_\_\ x4)$$

$$\_\_\ x4 = 3x\ \_\_$$

$$\_\_\_ = \_\_\_$$

$$(7x6)x3 = 7x(6x3)$$

$$\_\_\ x3 = 7x\ \_\_$$

$$\_\_\_ = \_\_\_$$

3. Where in materials were items presented?

   a. The format on this page differed from the test. The student was always required to write in product of the multiplication problems within the parentheses in the second line.

   b. The student also always had an arrow to aid him in putting the product in the correct place.

   c. This page also differed from the test in that the student solved each problem for both

equation types (axb)xc and ax(bxc). On the test he was required to solve only one side of each equation, eliminating a check of his work.

Once the evaluator has identified differences between the materials and the test, he must choose one possible cause of failure. If an inappropriate cause is selected, student performance will not improve and the evaluator will have to select another cause.

4. What caused the failure?

   Hypothesis to be tested:

   If the last page of the materials is changed to include problems similar to the test, then the student will pass the test.

5. How can the hypothesized cause of failure be tested?

   The following page (Figure III) was added as the last page in the materials. The student does not have arrows to indicate where the products are placed and he only answers one side of the equation.

EXAMPLE II - FIGURE III

Solve each equation:

$(2x5)x3 = 2x(5x3)$

$= 2x\bigcirc$

$= \_\_\_\_$

$(3x1)x2 = 3x(1x2)$

$= 3x\bigcirc$

$= \_\_\_\_$

$(2x7)x3 = 2x(7x3)$

$= 2x\bigcirc$

$= \_\_\_\_$

$(8x1)x3 = 8x(1x3)$

$= 8x\bigcirc$

$= \_\_\_\_$

$(3x5)x6 = 3x(5x6)$

$= 3x\bigcirc$

$= \_\_\_\_$

After this student completed this page, he passed an equivalent test. The revised last page was then included into the materials for all students. No further evaluation of these materials will occur until another student fails the same test.

There is no way for an evaluator to "know" if his hypothesized cause of failure is correct. If students have no further trouble with the materials and tests, the formative evaluation has improved the materials. If the same student or another student continues to have trouble, the formative

evaluation has not located the problem. In the first example, the addition of a new last page was sufficient for that student to achieve mastery of the objective. It was not for another student.

EXAMPLE II - FIGURE I

Another student fails the same test.

Write in the missing numbers using the associative principle.

A

(4x2)x5 = 4x(2x 5)

= 4x 10

= 40

B

(9x3)x6 = 9x(3x 6)

= 9 x 18

= 162

C

2x(4x8) = (2x4)x 8

= 8 x 32

= 256

D

6x(7x4) = (6x7)x 4

= 42x 28

= 70

After this student failed the test, the same procedures used in Example I were repeated. The student's test and materials were pulled for examination.

1. What was similar about the problems missed?
   a. Both problems missed were of the form (axb)xc.
   b. The student's errors on the second line were systematic. He multiplied (axb) and (bxc) in both problems he missed.
   c. The student's errors on the third line were different. In problem B he multiplied line 2,

in problem D he added line 2. The difference in the error on line 3 was considered of lesser importance by the evaluator because the student had not previously learned how to multiply two two-digit numbers, which could account for the addition.

d. Both items missed were on the right column on the page.

2. What was different about the items passed then failed?

   a. The items passed were of the form ax(bxc).

   b. Both items passed were on the left column of the paper.

3. Where were the items presented in the materials? For reasons similar to those discussed in Example I, the evaluator focused his attention on the last new page, presented in Example II, Figure II.

EXAMPLE II - FIGURE II

Solve each equation:

$(2x5)x3 = 2x(5x3)$

$= 2x\bigcirc$

$= \underline{\quad}$

$(3x1)x2 = 3x(1x2)$

$= 3x\bigcirc$

$= \underline{\quad}$

$(2x7)x3 = 2x(7x3)$

$= 2x\bigcirc$

$= \underline{\quad}$

$(8x1)x3 = 8x(1x3)$

$= 8x\bigcirc$

$= \underline{\quad}$

$(3x5)x6 = 3x(5x6)$

$= 3x\bigcirc$

$= \underline{\quad}$

a. The page was done correctly by the student.

b. All the problems on the page were of the form ax(bxc).

4. What caused the failure?

Hypothesized cause of failure:

If the last page of the materials is revised to include practice in doing both forms (axb)xc and ax(bxc) of the associative principle, then the student will pass the test.

5. How can the hypothesized cause of failure be tested?

The last page of the materials was again revised to include problems of both forms ax(bxc) and (axb)xc of the associative principle.

Once this page was revised, this student was given an equivalent form of the test. He and his fellow students had no further trouble with this objective during the year.

These two examples were chosen to illustrate not only how to use this model but to demonstrate how the use of student test performance can be used to continually evaluate the evaluator's decisions. No original instructional materials or revision are ever free from revisions.

EXAMPLE III - FIGURE I

Skip count by 3's.

472,	475,	478,	471,	474,	477,	490	X
205,	202,	199,	196,	193,	190,	187	X
747,	750,	753,	756,	759,	762,	765	
1,000,	997,	994,	991,	988,	985,	982	X
638,	641,	644,	647,	650,	653,	656	

The third example has been selected to illustrate how procedures can identify a wide range of causes of failure. This illustration demonstrates why an evaluator must consider not only current instructional materials but also the prerequisite materials. In Example III, Figure I is a student's test for an objective requiring the student to skip count by 3's.

1. What is similar about items failed?
   a. The pupil's errors in skip-counting are always made when the pupil has to change the place value in the tens or hundreds place.
2. What is different about items passed than failed?
   a. The pupil can skip-count by 3's when the place value does not change.
   b. The pupil can skip-count by 3's and change place values for some multiples of 10 but not for all.
3. Where in the materials is the content presented?

   In looking over the student's materials, the pupil consistently made errors when he had to change place values and there were no clues about the value of the new place value. Since the materials were designed to teach skip-counting by 3's and the student did demonstrate he could skip-count without changing place values, the evaluator decided to look at the prerequisite behaviors. The immediate prerequisite behavior required that the student be able to count by 1's. The criterion test is presented in Example 3, Figure II.

EXAMPLE III - FIGURE I

Count Forward by 1's:

A 374, 375, 376, 377, 378, 379, 380, 381 X

B 995, 996, 997, 998, 999, 1000

C 230, 231, 232, 233, 234, 235, 236

D 659, 660, 661, 662, 663, 664, 665, 666

Count Backward by 1's

E 529, 528, 527, 526, 525, 524, 523

F 837, 836, 835, 834, 833, 832, 831

G 311, 310, 309, 308, 307, 306, 305

Since the student's failed test indicated a possible problem in counting by 1's, the test for this behavior was examined. The series in line A was the only series which required the student to name the next place value without any clues except for line B which was considered because of prior experience, easier for students to learn. Line D and G, although they required a change in place value, also provided clues further in the series as to what the new value should be. If a student only missed line A, it would be possible to be given mastery if the evaluator, at that time, was not aware of the uniqueness of this line.

4. What caused the failure?

Hypothesized cause:

If a student cannot count by 1's to 1000, then he cannot skip count by 3's correctly.

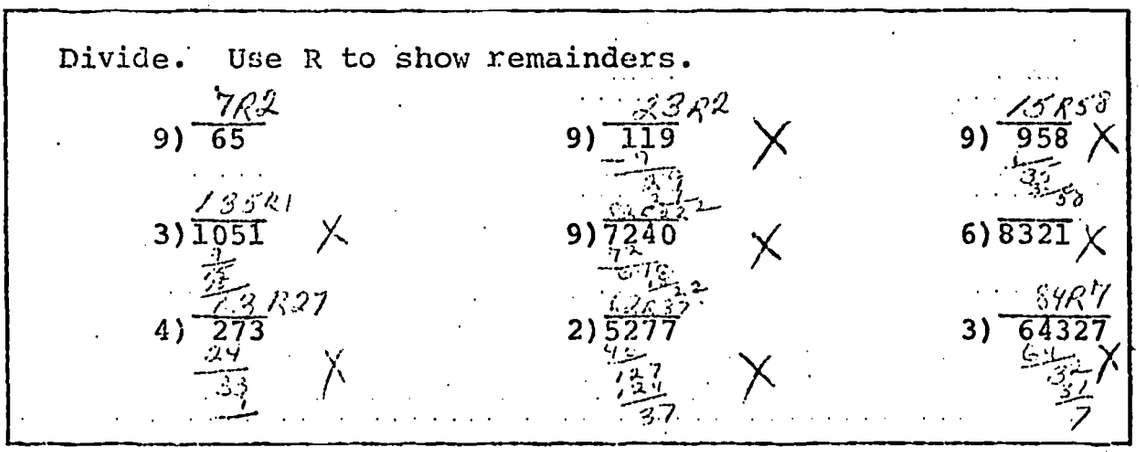
5. How can the hypothesized cause of failure be tested? A revised test was constructed to include more place value changes without providing clues about the new place value. This student was not able to master this test so he was reassigned the materials to teach him how to count to 1000.

The revised test was substituted into the curriculum. No student who mastered the revised test failed the test of skip counting by 3's.

This example illustrates how lack of prerequisites can cause inadequate performance in future objectives. The evaluator must consider all aspects of an instructional system which can effect student performance not only the failed test and its associated materials.

When a pupil fails a test, there may not be anything wrong with the instructional materials but the problem could be the way in which they were used by the student. Our instructional materials were designed to be used in a specified way. Evidence indicated that deviations from the specified procedures could result in inadequate test performance for some pupils. The evaluators always had to consider inappropriate work skills as a potential cause of failure which could be tested. An example of this can be found in Example IV, Figure I.

EXAMPLE IV, FIGURE I



1. What was similar about problems failed?

   a. The errors do not appear to be systematic.

   1. 9) 119 (23 R2)
      9
      29
      27
      2

      The student put the remainder of 2 as the first digit in the quotient or he multiplied 9x2=9. He finished the problem correctly.

   2. 3) 1051 (35 R1)
      9
      15
      15
      1

      The student was correct until he got to the third digit in the quotient, where he left out the 0 before stating the remainder.

   3. 2) 5277 (62 R37)
      40
      127
      124
      37

      There does not appear to be any reason for his responses to this problem.

   b. The student appears to think all answers must be 2-digit numbers with a remainder.

   c. There were many erasures on the test which could indicate confusion.

2. What is different about items passed than failed?

a. The only problem done correctly was a two-digit number.

3. Where was the content presented?

The student's work pages were examined. The last page of the materials was done perfectly by the student. The page contains quotient with more than two digits and the pupil never had a remainder greater than the divisor.

Students in our classes were allowed to score their own work pages. They had access at all times to the answers to their work pages. Since this student only answered one problem correctly on the test, his errors were inconsistent, and his work pages were perfect, the evaluator had to consider the possibility that the student misused the answer keys.

4. Why did the student fail?

Hypothesis:

If a student uses an answer key to do his instructional materials with, then he will fail the test.

5. How can the hypothesized cause of failure be tested?

The student was reassigned the identical materials except the teacher scored his work pages. If the student had not used the key incorrectly, he should be able to do the pages correctly again and still fail the test. If he had misused the key, he will no longer have a key to answer his pages with. If

If the materials are faulty, he will still fail the equivalent test. If the materials are adequate, he should pass the test.

The student had difficulty in re-doing his materials. After he completed the assigned materials, he passed the equivalent test. Since no other student during the year failed the test, the evaluators felt their hypothesis of poor work skills was correct.

The preceding examples have been chosen to illustrate how these procedures for formative evaluation are used to detect inadequate materials, hypothesize a probable cause, and test the proposed hypothesis. The successful use of these procedures is heavily dependent on the evaluator's skill in analyzing individual test items. It is difficult to describe how one can examine a test and find similarities and differences between the items passed and failed. Practical experience appears to be the best and perhaps only teacher. The more tests that are examined by the evaluator, the easier it becomes to identify differences and similarities. Also, because the procedures provide the evaluator with an immediate evaluation of his own skills by requiring an immediate test of all hypotheses, skills in detecting errors can be continually improved.

The examples have demonstrated how poor programmed instructions, lack of prerequisites, and inappropriate student study skills can lead to inadequate test performance. There are many other causes of poor test performance which can be

identified and tested using these procedures. Some others include lack of student motivation, non-equivalent tests, omission of prerequisites, poor criterion test mastery, insufficient practice pages, and omission of teaching unique type items or items which are exceptions to a rule.

The more difficult problem for the evaluator is in defining a probable cause of test failure which is testable. It is easier to decide what a student has not mastered than to decide why he has not learned something from the materials. Knowing a student failed subtraction problems which entailed borrowing is not the same as hypothesizing why he did not learn to borrow from the materials. To translate the what into why, the evaluator has to gather information about if it was taught, if it was learned, if the pages were done correctly by the student, if the correct pages were done by the student, and if the pages were scored correctly by the student.

In order to decrease the number of hypotheses which have to be tested for any one student, a category system for generalizing types of errors to usual causes was created. After examining many tests, certain relationships between types of incorrect student answers and usual causes became apparent.

One crucial relationship was the number of test items failed by a student. When a student missed one or two problems on a test, the cause was more likely to be created by a unique difference in problem content between the test item and instructional pages.

When a student missed almost all test items, the errors were then analyzed in terms of being random or systematic and computational or process. Systematic errors consist of using the identical rule to do all the problems incorrectly. An example of a systematic error can be found in Example I. The student always wrote the product within both sets of parentheses as the answer, irrelevant of the question. Systematic errors usually indicate a discrepancy between what is taught in the materials and how it is tested.

Random errors consist of doing problems incorrectly for different reasons, such as in Example IV. These types of errors usually result from a misuse of the materials by the student.

Computational errors consist of setting up the problem correctly but when either adding, subtracting, multiplying, or dividing, writing in the incorrect answer. These types of errors are often considered careless errors. Students who fail an objective because they do not know their number facts are failing because they lack a prerequisite skill. These types of errors cannot usually be corrected by revising the lessons, but require practice in the prerequisites.

The final type of error is a process error where the student fails because he has not learned the concept or process being taught. For example, he cannot set up a multiplication problem, but instead adds the two numbers. Failures resulting from this type of problem usually indicate poor study skills or inadequate teaching pages.

These four categories can overlap and when they do, it can confound the problem for the evaluator. The model restricts changing one dimension of the materials at a time. The evaluator must decide which probable hypothesized cause of failure should be tested first. It is preferable to test student study skills first if there is supporting evidence because this allowed a better evaluation of the existing materials; but if evidence in the booklet indicates study skills are adequate, the evaluator must search for the cause within the materials. If a hypothesized cause of failure is not obvious after deciding what is similar and different about items passed and failed on the tests, the evaluator can use other clues in the materials to decide why the student has not learned. In looking over the student's materials, the evaluator can learn where the student's trouble started. This information can be gotten from looking for the first page with many errors, many erasure, or messy pages. If these clues appear in the first few pages, the evaluator may consider the lack of prerequisites as a possible cause. If these clues appear after teaching pages, the evaluator should consider how well did the pages teach or how attentive was the student in class.

Each person using the model will discover his own clue system. What the model does offer is a way to test quickly and objectively each proposed hypothesis in order to accept or reject it. Because the evaluator must find a way to improve each student's performance, all problems the students have

with the materials are tracked down. The results of this type of evaluation should be immediate improvement in materials as they are being developed.

The systematic use of these formative evaluation procedures appears successful in improving instructional materials. The use of a design based on eliminating rival hypotheses was effective in identifying causes of test failure and in identifying variables which effect academic performance. Because these procedures require the evaluator to establish cause and effect relationships by disproving alternative explanations, he is forced to consider all variables as possible explanations of inadequate pupil performance. To decrease the number of probable alternate hypotheses, it is suggested that the evaluator remove the effects of some variables in the classroom by either eliminating their influence or by keeping their influence on academic performance consistent.

Inadequate test performance was contributable to two major sources: inadequate instructional materials or inadequate student study skills. When any cause of test performance was attributed to student study skills, instructional materials were not revised. Revisions in materials were made after one student's performance indicated an inadequacy in the instructional materials. This use of one student's performance to evaluate any set of materials has two advantages. It allows the evaluator to quickly identify and improve materials because once a cause is located by one student, revisions can be made

immediately. The other advantage is other students can be used to evaluate the revisions in a continuing process.

The use of these procedures have many advantages for the curriculum writer and evaluator. These procedures were very effective in identifying errors attributed to all components of the instructional system including poorly constructed tests, and inappropriate student performance. All causes of test failure must be identified by selecting and eliminating rival hypotheses. By requiring all students to pass an equivalent test before a cause and effect relationship can be established, the evaluator is forced to systematically identify and consider all aspects of a classroom which can effect academic performance.

The most promising outcome of this study was that systematic formative evaluation during the in-context tryout of materials is feasible. Although curriculum designers cannot use classical experimental designs in evaluating materials, other designs appear practical. Systematically eliminating rival hypotheses appears useful in identifying inadequacies within an instructional system and in generating appropriate revisions. Because revisions are made quickly, the instructional materials are improved and tested during the development of the materials.

References

Campbell, D. T. From description to experimentation: interpreting trends as quasi-experiments. In C. W. Harris (Ed.) Problems in measuring change. Wisconsin: University of Wisconsin Press, 1965.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally and Company, 1963.

Cronbach, L. J. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-83.

Hastings, T. J. Curriculum evaluation: the why of outcomes. Journal of Educational Measurement, 1966, 3, 27-32.

Lumsdaine, A. A. Assessing the effectiveness of educational programs. In R. Glaser (Ed.) Teaching machines and programmed learning II, DAVI, NEA, 1965.

Merkle, S. M. Good frames and bad. New York: John Wiley and Sons, Inc., 1964.

Platt, J. R. Strong inference. Science, 1964, 146, No. 3642, 347-353.

Sidman, M. Tactics of scientific research. New York: Basic Books, Inc., 1960.

Stufflebeam, D. L. The use of experimental design in educational evaluation. Paper presented at the national convention of the American Educational Research Association, Minneapolis, Minnesota, 1970.