DOCUMENT RESUME

ED 090 255 TM 003 528

AUTHOR Moncrief, Michael H.
TITLE Procedures for Empirical Determination of En-Route Criterion Levels.
PUB DATE [74]
NOTE 45p.; Paper presented at the American Educational Research Association Annual Meeting (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS Decision Making; *Educational Diagnosis; Educational Objectives; Elementary School Mathematics; Individualized Instruction; Instructional Systems; Instructional Technology; *Learning Readiness; Management Systems; Performance Criteria; Prediction; *Sequential Learning; Student Evaluation; Test Validity

ABSTRACT

En-route Criterion Levels (ECLs) are defined as decision rules for predicting pupil readiness to advance through an instructional sequence. This study investigated the validity of present ELCs in an individualized mathematics program and tested procedures for empirically determining optimal ECLs. Retest scores and subsequent progress were validating criteria. Results indicated empirical data can identify more efficient ECLs than those established a priori. To justify the cost involved, however, such data should be collected by the instructional designers during the formative evaluation field test of a program, using automated data processing and multiple-matrix sampling techniques. (Author)

PROCEDURES FOR EMPIRICAL DETERMINATION OF EN-ROUTE CRITERION LEVELS

Michael H. Moncrief
SWRL Educational Research and Development

ABSTRACT

En-route Criterion Levels (ECLs) are defined as decision rules for predicting pupil readiness to advance through an instructional sequence. This study investigated the validity of present ELCs in an individualized mathematics program and tested procedures for empirically determining optimal ECLs. Retest scores and subsequent progress were validating criteria. Results indicated empirical data can identify more efficient ECLs than those established a priori. To justify the cost involved, however, such data should be collected by the instructional designers during the formative evaluation field test of a program, using automated data processing and multiple-matrix sampling techniques.

Paper presented at the 1974 AERA Annual Meeting (Session 28.09)

PROCEDURES FOR EMPIRICAL DETERMINATION OF EN-ROUTE CRITERION LEVELS

Michael H. Moncrief

Recently there has been increased discussion among educators about designing instruction for mastery learning. This discussion has often revealed a lack of clarity as to the meaning of mastery.

In an educational context mastery of a task generally is considered to have been attained (a) when an individual demonstrates sufficient proficiency to perform a given function, or (b) when he can perform well enough on that task to benefit from being advanced to the next.

Using the first criterion, it seems reasonable to say that a student has "mastered" a given instructional objective when he can be expected to perform satisfactorily in those situations typically found in his everyday life that call for the use of that which has been learned. This means he must be able to retain and apply some minimal amount of what he has learned.

The other educational criterion used in defining mastery is the one most frequently used and is much easier to determine empirically. As Glaser (1963) suggests, "mastery" can be used to specify the minimum proficiency the student needs to demonstrate before going on to the next instructional unit in a sequence. Using this definition the most efficient operationalization of mastery would seem to be that proficiency level which maximizes the subsequent progress of students through the instructional units.

Currently available individualized instructional programs have specified their mastery levels, and thus their instructional management

decision criteria, mainly on the basis of intuitive judgment. Some empirical data are needed for use in establishing "mastery levels" which maximize the efficiency of selected instructional management decisions.

This study investigated the validity, as instructional management decision rules, of the preset en-route criterion levels (ECL) associated with a selected group of instructional objectives in an individualized mathematics program. The validation was conducted in terms of delayed retest scores and in terms of subsequent progress through the instructional continuum. The empirical data gathered in this study were used to suggest optimal performance standards for use with the selected objectives. The study also examined the cost/benefits associated with the use of ECLs derived by empirical procedures.

Explanation of Terms

There are many tasks involved in managing an individualized instructional program. This study, however, was directed only at a very specific subset of those tasks. The specific management tasks of interest were those concerned with directing student progress through a continuous individualized instructional program. In this study the term "instructional management" is used to refer to these management tasks. "Instructional management decisions" refer to those decisions, made on the basis of ECLs, which determine whether or not a student is ready to advance to the next learning task in a sequence.

In the context of most individualized instructional programs, including the one studied here, instructional management decisions are

made on the basis of criterion-referenced tests. These tests are designed to measure a student's performance in a prespecified domain that has been operationally defined. The score on a criterion-referenced test is thought to indicate the degree of proficiency that a student has attained on that specific objective. In order to proceed from one learning objective to the next, a student must demonstrate a certain degree of proficiency on the first objective, i.e., he must meet or surpass the performance standards or criterion level for that objective. Such proficiency levels are often referred to as "mastery" levels and a student who scores at that level or above is said to have attained mastery.

The use of mastery levels is found in many of the recently developed instructional programs. For instruction using mastery levels the typical paradigm, illustrated in Figure 1, requires the student to attain a predetermined criterion score on each objective or unit of instruction before advancing to the next unit in the sequence. Usually any student who fails to reach the criterion is recycled through that segment of the program, often receiving some form of remedial instruction. Recycling is continued until the desired proficiency is demonstrated. This is the procedure followed by most of the current individualized instructional programs, including Individually Prescribed Instruction (Cooley and Glaser, 1969), Project PLAN - Program of Learning in Accordance with Needs (Flanagan, 1968) and Individualized Mathematics System (Ironside, 1971). It is also the paradigm used in most Computer Managed Instruction projects (Lawler, 1971).

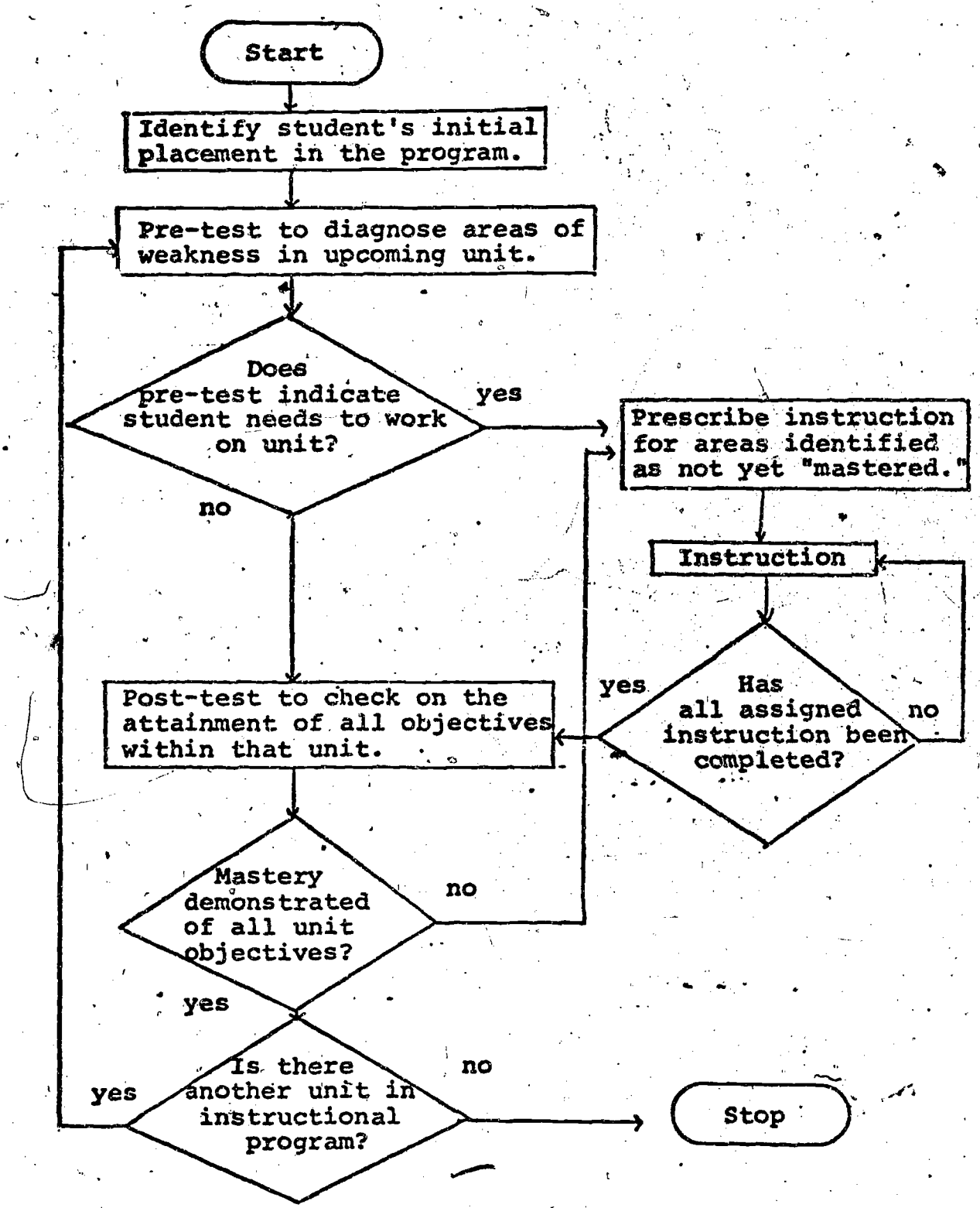

Figure 1.--General Paradigm of the Instructional Management Process in Individualized Instructional Programs

There are some very real costs involved in procedures requiring the use of mastery tests, but there are even greater costs involved in any incorrect instructional decisions made on the basis of these tests. In most individualized programs, for example, a student who scores just below the required level is recycled through that instructional unit. This recycling could be considered unnecessary if, had he been advanced, his subsequent progress through the program essentially would be unchanged. This procedure incurs unjustified costs to the system in terms of instructional time, the unnecessary use of materials, limiting the progress of the student, and a decrease in his motivational level. Much the same cost also would be involved if a student were allowed to advance before he had sufficient proficiency in the prerequisite skills needed for a reasonable probability of success on the subsequent unit.

To date, mastery levels have been set on an *a priori* basis. To optimize the probability of making correct instructional decisions an empirical basis for establishing criterion scores is needed. If empirical procedures are found to produce decision points which are significantly different from those established on an intuitive or purely "theoretical" basis then the application of those procedures could make a substantial improvement in the operation of individualized instructional programs. Such procedures applied to appropriate instructional programs could be expected to facilitate the progress of students through those instructional sequences and thus reduce instructional costs. Whether or not this reduction represents a substantial savings largely depends on the costs of applying the empirical procedures. The current practice in many

instructional programs is to establish the same criterion levels for all students over all objectives. Some evidence is needed pertaining to the soundness of requiring uniform levels of proficiency over all objectives and over all students. As the number of individualized programs of instruction being developed and implemented increases, so does the need to come to grips with these instructional design questions.

Criterion-Referenced Tests and Instructional Management

The question of the appropriate uses of criterion-referenced and of norm-referenced measures is still somewhat unsettled. Convincing arguments, however, have been presented for the use of criterion-referenced assessment procedures in making instructional management decisions (Glaser, 1963; Gagné, 1965; Popham and Husek, 1969). Criterion-referenced tests are designed to provide information about the degree of competency attained by individual students irrespective of the performance of others. The use of such tests is required in programs where "mastery learning" is the goal (Airasian, 1971), in individualized programs (Coulsen and Cogswell, 1965) and especially in the management of such instructional programs (Kriewall, 1969).

The Establishment of Criterion Levels

Performance levels have been used in education for a long time, but their use in connection with criterion-referenced tests was mainly fostered by the increasing popularity of programmed instruction and Mager's (1962) work on instructional objectives.

Since Bloom (1968) popularized the term "mastery learning," it generally has been agreed that the performance requirements should

represent "mastery" of the objective being learned. However, the particular criterion levels used in any one program are established, at best, on the basis of experienced judgment and intuition.

Although a number of writers suggest that they are designing instructional programs around the idea of mastery learning, there is great variation in the way they have operationalized mastery. A recent survey of a number of instructional programs indicates a tendency to specify a rigid criterion selection policy within a given program. It also was reportedly not difficult to find programs "where higher criteria are selected in the mistaken belief that this will result in a better quality of learning product than will a system having a lower criterion" (Kriewall, 1969, p. 52).

The lack of any specific agreement as to the operational meaning of mastery learning also can be seen in the way criterion levels have been established by those who purport to be developing "mastery learning" in their instructional programs. Mager suggests that one way to determine "how excellent (a student) must be before we will consider him satisfactory ... is to look over the examinations you use. They will tell you what you are using as standards of performance ..." (Mager, 1962, pp. 51-52). Bloom (1968) reports having used a procedure similar to that outlined by Mager. He established his criterion points for one class on the basis of the standards used in grading the students in the previous year's classes.

As Block (1971) points out, at present there are no established rules for setting mastery standards and, until there are, instructional designers must rely on procedures similar to that used by Bloom or set

the standards on a purely subjective basis. However, the use of each of these procedures has yielded quite varied results. In a program designed to teach reading skills, mastery has been defined as 95% proficiency (Hackett, 1971). In contrast, the Individually Prescribed Instruction (IPI) project has generally operationalized mastery as 85% proficiency (Bolvin, Lindvall, and Scanlon, 1967; Glaser, 1968). The Individualized Mathematics System (IMS) has defined mastery in terms of various performance levels depending on the number of test items and the importance of the objectives to future learning (Ironside, 1971). Although the IMS test authors were asked to establish mastery scores above the 75% level, in practice IMS has a range of proficiency levels from 66% to 100% with the most typical level being about 80%.

Thus far criterion standards have been variously set at admittedly arbitrary levels (Bloom, 1968; Kriewall, 1969; Merrill, 1971). There has been no empirically verified procedure for establishing the criteria upon which to base the type of instructional management decisions considered in this study.

The Effectiveness of Mastery Learning Strategies

Most of the research conducted in the area of mastery learning has dealt with the general overall effectiveness of the strategy. Such research has been concerned with testing the hypothesis that it is advantageous to have students demonstrate the attainment of intuitively established minimal proficiency levels. There has been no research, with the possible exception of Block's (1970), designed to determine optimal proficiency levels or the method of arriving at them.

Mastery learning strategies generally have been found to be successful in terms of both cognitive and affective results (Postlewait, Novak and Murray, 1964; Biehler, 1970) especially when success is defined in terms of the percentage of students attaining previous grading standards and receiving top grades (Mayo, Hunt and Tremmel, 1968; Kersh, 1970). Even though there is no empirical basis for whatever particular performance standards are used, there is evidence that greater learning is achieved by students required to attain some criterion performance than by students for whom no requirements are made (Block, 1970; Lawler, 1971).

Need for Empirical Evidence in Defining Criterion Levels

The fact that criterion levels have been established more on the basis of the intuition of the individual instructional designers than on the basis of empirical evidence does not mean that the instructional designers are insensitive to the need for such evidence. On the contrary, many of the writers as well as other concerned educators and psychologists have discussed the basis upon which criterion levels should be established. Glaser (1963) has suggested that we need to specify the minimum proficiency levels the student needs before going on to the next instructional unit in a sequence. Although the difficulties involved prevented them from doing so, Bloom, Hastings and Madaus (1971) note the desirability of using carefully worked-out performance standards. They also discuss the important relationship between "appropriate mastery levels" and student motivation.

One discussion which summarizes part of the need for empirical evidence related to performance criteria is presented in a manual for

the IPI Institute.

> The determination of specific mastery levels for various subject matter is an experimental problem which needs to be studied. How much mastery should be required, for example, in learning basic arithmetic facts before moving on to an advanced topic? Is more rapid learning and better retention achieved if a student is permitted to go on in a subject matter where advanced lessons depend on previous lessons or is it best to require an early high level of mastery? (In teaching typing, for example, it may be best to permit the beginning typist to make errors without compromising her speed so that eventually both speed and accuracy are learned efficiently.) (Bolvin, Lindvall and Scanlon, 1967, p. 8).

Mastery learning requirements, though only intuitively established, generally have been found to be effective. It seems reasonable, then, that an empirical investigation of the most efficient operational definition of mastery for particular instructional programs could be expected to increase further the effectiveness and efficiency of such programs.

Realizing the need for some evidence to help establish the criterion scores used in mastery testing situations, Block (1970) investigated the effects, both cognitive and affective, of requiring the attainment of various pre-established "mastery" levels. In that study, ninety-one eighth grade mathematics students were presented with a one week instructional program on matrix algebra. The instruction was presented in three sequential units. Students were randomly assigned to different groups which were required to attain and maintain selected mastery levels (no requirement, 65%, 75%, 85%, and 95%) on each unit as they advanced through the program. Block found that the performance of each of the mastery groups was greater than that of the group for which no criterion level was required. He also found that while

the highest mastery level produced the greatest cognitive learning, it had a negative effect on student interests and attitudes, suggesting that a somewhat lower criterion level would be more advantageous in terms of the balance between cognitive and affective results.

The Block investigation differs from the present study in a few important ways. First, the Block study was conducted in the context of a relatively brief and self-contained instructional program. Second, the students were not only required to attain a given degree of proficiency on each unit, but also to maintain that same degree of proficiency as they advanced from one unit to the next. This is not the typical paradigm utilized in most instructional programs. Most important, the Block study viewed the criterion levels as treatment variables rather than decision rules. Thus, there are corresponding methodological differences.

The Process of Validating Instructional Decisions

To maximize the efficiency of instructional management decisions some evidence is needed as to the proficiency required on a given objective before a student can be expected to continue successfully through the instructional program and retain the material that has been learned. Basically this amounts to the establishment of mastery levels which will be used as decision rules, to predict future performance. In using mastery tests instructional designers should be concerned not only with the content validity of the tests, but also with the validity of the decision rules represented by the criterion scores.

Cronbach (1970) suggests that the "validation of a decision rule logically requires an experiment in which after being tested, persons are allocated to treatments without regard to the scores whose usefulness is being validated. The outcomes of the treatment are then appraised." Cronbach also suggests that the emphasis not be on a validity coefficient but rather on the relationship between the outcome measure(s) and the test score. This procedure is valid only insofar as the validating criterion is truly representative of the outcome we wish to measure.

METHODOLOGY

The specific questions asked in the present study were:

(1) How valid, as instructional management decision rules, are the preset performance criteria associated with a selected group of instructional objectives (a) in terms of subsequent progress through an individualized instructional continuum, and (b) in terms of delayed retest scores?

(2) On the basis of the data gathered in this study, what is the apparent optimal performance criterion for use with each of the selected objectives?

(3) On the basis of this study what, if any, are the probable cost/benefits which could be expected from the use of criterion levels derived by empirical procedures?

Curriculum Context

To obtain data which reflect the effects of [redacted] independent variable, performance criteria, a number of requirements were imposed upon the

curriculum context in which this study was conducted. It was critical that the study be conducted within the context of an instructional program which makes instructional management decisions on the basis of "mastery" scores. The program also should have been field tested to insure that it was operational and generally effective.

The program selected for this study was the Individualized Mathematics System (IMS). This program was developed by the National Laboratory for Higher Education and field tested during the 1969-70 and 1970-71 school years (Frary, 1971). IMS uses a variety of instructional techniques to teach the many objectives found in its purported hierarchically and logically sequenced continuum. As in most individualized programs, student progress is controlled by the decisions made on the basis of the posttests administered at the end of each instructional unit.

In the IMS program each unit is composed of a number of objectives. Although the posttests for all unit objectives are found at the end of that unit, the advancement-recycle decision is made on the basis of each separate objective posttests. Thus, after taking a unit posttests a student is recycled through that part of the instructional program related to those objectives, and only those objectives, for which he failed to achieve a "mastery" score.

Operational Definitions

For the purposes of this study the students' postinstructional proficiency levels were measured by the posttests supplied by the IMS instructional program. Retention was measured by the use of parallel tests developed using the statement of the objective and the existing tests of that objective as guides. An indication of a student's

subsequent progress through the program was obtained from his posttest score on each objective in the next sequential unit of instruction.

Setting

This study was conducted during the last half of the 1971-72 school year in the American Elementary School in Karlsruhe, Germany, as part of that school's overall attempt to individualize its instructional program. The Karlsruhe school had an enrollment of about 1,000 pupils. The children came from a variety of ethnic and economic backgrounds. Most were the dependents of Army personnel, both military and civilian. Being the only American school in Karlsruhe, it also served as the educational institution for the children of most of the American businessmen residing in the area.

Subjects

The subjects used in this study were selected from among those pupils at Karlsruhe who were working on a specified subset of IMS objectives during the time the study was being conducted. It seems reasonable to suspect that the sample of pupils working in these units during this given time period were representative of the Karlsruhe pupils for whom the units were appropriate. A further description of the selection of the subjects is presented in the next section.

Research Design and Data Gathering Procedures

This study was designed to investigate the validity of the go/no-go decision rules associated with the posttests found at the end of a selected set of instructional units in a sequential individualized program.

The general procedures followed in this study were those suggested by Cronbach (1970) for the validation of decision rules. To study the accuracy of the decision rules associated with a particular unit, the subjects who took the posttest for that unit were advanced to the next sequential unit without regard to their posttest scores on the first unit. For convenience these units are designated here as A and B, respectively, with subscripts indicating the different pairs involved in the study. A retest of the unit A objectives was administered as soon after the unit B posttest as possible. The data gathering design for each pair of units followed this order:

$$\frac{\text{Unit A}}{\text{Instruction}\rightarrow\text{Posttest}} \rightarrow \frac{\text{Unit B}}{\text{Instruction}\rightarrow\text{Posttest}} \rightarrow \frac{\text{Unit A}}{\text{Retest}}$$

Figure 2 shows the units within the IMS continuum which were selected for study. As shown in Figure 2, the IMS continuum consists of ninety units organized by topics and levels of difficulty. Each unit contains from one to eleven objectives. Generally, a pupil moves through this sequence in order, that is, from top to bottom and left to right, each unit purportedly building upon the preceding units.

The four program segments shown in Figure 2 were selected on the basis of the identification of the concentration of dependable student data, the availability of instructional materials and a growing awareness of the difficulties involved in the collection of dependable data.

Implementation of the described research procedures was initiated on January 18, 1972, with the actual collection of data beginning on

IMS Topics	Levels								
	I	II	III	IV	V	VI	VII	VIII	IX
Numeration					A_3				
Addition					B_3				
Subtraction									
Multiplication					A_4				
Division					B_4				
Fractions				A_1					
Mixed Operations				B_1					
Money									
Time				A_2					
Systems of Measurement				B_2					

Figure 2.--Schematic representation of the selected units within the IMS continuum.

February 7, 1972. Data collection was completed on April 17th of that year. The data consisted of posttest scores for each objective in the first unit (A) of each pair, posttest scores for each objective in the second unit (B), and retest scores on the objectives in the A units obtained from carefully constructed alternate forms of the respective posttests. Table 1 shows the number of objectives included in each of the units studied.

Ten classrooms were involved in this study. They consisted of all of the 4th, 5th, and 6th grade classes at Karlsruhe except for those few in which the IMS program had not been sufficiently implemented to provide a truly representative IMS classroom situation.

Data were collected on all pupils who were: (1) in one of the classrooms where data was being collected, (2) working on the specified units during the time the study was being conducted, and (3) who, upon completing a given unit A, were to advance to the next sequential unit in the program.

The teachers and paraprofessionals at Karlsruhe were directly involved in implementing the design procedures and especially the data collection activities required in this study. The teachers were asked to make certain after a subject completed a Unit A posttest he did not do any further work in any IMS material for that unit. However, the subjects were exposed to the ongoing instructional programs presented in their respective classrooms. At times this included exposure to material on topics related to those which they had just studied. After completing a Unit B posttest, each subject was given the appropriate retest. The teachers and paraprofessionals were asked

Table 1

UNIT PAIRS ON WHICH DATA WERE COLLECTED

Pair Number	Unit A	Number of Objectives	Unit B	Number of Objectives
1	Level IV Fractions	5	Level IV Mixed Operations	4
2	Level IV Time	5	Level IV Measurement	4
3	Level V Numeration	4	Level V Addition	8
4	Level V Multiplication	6	Level V Division	5

to administer the retest as soon after the Unit B posttest as possible. Often this was either the same day or the next. Due to the record keeping procedures in a few classrooms it was sometimes two or three days before the retest was taken by a subject. Due to the continuous monitoring of classroom activities, testing procedures, and IMS records, it was never longer than one week between Unit B posttests and the associated retests.

ANALYTICAL PROCEDURES AND RESULTS

Validation of Program Performance Requirements and Identification of Optimal Criterion Levels

The decision rules investigated in this study were those associated with each objective found in the selected four IMS units; Fractions IV, Time IV, Numeration V, and Multiplication V. Data were gathered on a different group of pupils for each of these four units. Table 2 shows the number of subjects for whom data were obtained for each unit. The four selected units contained a total of twenty objectives. Thus, there were twenty decision rules whose validity were investigated in the present study. For convenience, these objectives are often referred to by their respective reference numbers, as presented in Table 2.

In analyzing the validity of these decision rules the scores on each objective posttest were dichotomized at all possible decision points. For example, where an eight-item posttest was involved, the subjects were classified as go or no-go on the basis of those who scored 8, 7 or higher, 6 or higher and so forth. Each of these dichotomized groups, which represented different decision rules, were cross tabulated with

Table 2

REFERENCE NUMBERS, CORRESPONDING IMS DESIGNATION, NUMBER OF TEST ITEMS FOR EACH OBJECTIVE, AND NUMBER OF SUBJECTS PER OBJECTIVE

Assigned Reference Number	IMS Designation		Number of Associated Test Items	Number of Subj. for Whom Data Were Collected
	Unit	Objective Number		
A-1	Fractions Level IV	1	8	N = 62
A-2		2	9	
A-3		3	8	
A-4		4	11	
A-5		5	7	
A-6	Time Level IV	1	3	N = 49
A-7		2	6	
A-8		3	4	
A-9		4	4	
A-10		5	6	
A-11	Numeration Level V	1	6	N = 31
A-12		2	4	
A-13		3	7	
A-14		4	13	
A-15	Multiplication Level V	1	6	N = 26
A-16		2	6	
A-17		3	23	
A-18		4	8	
A-19		5	8	
A-20		6	6	

each of the dichotomized validating criteria to form a series of 2 X 2 contingency tables and a contingency coefficient was computed for each. These contingency coefficients were then used to judge the validity of the IMS program decision rules and to identify the apparently optimal decision point or mastery level associated with each selected objective.

For clarity, the meaning of the terms performance standards, criterion level and mastery level, which are used interchangeably, need to be explained in the context of the present study. Most of the discussions about mastery levels define them in terms of some percentage score. In many instances these definitions or descriptions of the required performance levels are misleading, at best, when translated into actual program operating procedures. For example, suppose an instructional program purports to be requiring a 95% criterion level, but is using tests of less than twenty items. Then that program, in actual practice, is either requiring a criterion level of 100% or some performance which might be substantially below 95%.

To avoid discrepancies and because of the variation in the length of many IMS posttests, performance standards are reported for this study in terms of the number of correct item responses required. Thus, when the program performance criterion for a given objective is reported as being level 7, it means that a pupil has to get at least 7 items correct on that post-test before he would be advanced to the next unit. Similarly, a suggested optimal criterion level of 6 on that test means that, according to the data obtained in the present investigation, it apparently would be maximally efficient to base the advancement decisions

associated with that objective on whether or not a score of 6 or higher had been attained. To make these performance levels meaningful and interpretable, the number of items associated with each objective is given in Table 2.

As represented in Figure 3, two criteria were used in selecting the optimal decision point associated with each objective, (a) there must have been a significant ($\alpha \leq .10$) contingency coefficient between scores dichotimized at that point and at least one of the validating criteria; and (b) the number and magnitude of significant contingency coefficients between it and all validating criteria must have been maximal, i.e., greater than any set of significant contingency coefficients associated with any other decision point.

If, and only if, the IMS decision point was optimal as defined above, was it designated valid as a decision rule. If the analyses for a particular objective produced an optimal decision point for that objective which was other than the one given in the IMS program, the program decision rule was designated invalid, and the optimal decision point was designated as the valid decision rule.

If the analyses for any given objective yielded two or more decision points the number and magnitude of whose respective significant contingency coefficients were approximately equal, and if the IMS decision point was among them, it was designated valid as a decision rule. If, however, the IMS decision point was not among the contending points, that contending point which differed least from the IMS decision

	Maximal Contingency Coefficients	Not Maximal Contingency Coefficients
SIGNIFICANT ($p \leq .10$ on at least one criterion)	OPTIMAL	Not Optimal
NOT SIGNIFICANT ($p > .10$ on each of all criteria)	Not Optimal	Not Optimal

Figure 3.--Schematic representation of an optimal decision point.

point was designated the optimal and valid decision rule. This procedure minimizes the number of students affected by a change in mastery levels, when that change is between competing decision points. It thus yields the more conservative estimate of the cost/benefits discussed in section three of this chapter. This procedure also lends some weight to the intuitive knowledge of the instructional designers as reflected in the a priori program mastery level.

If the analyses for a given objective produced no statistically significant ($\alpha \leq .10$) contingency coefficient between any of its possible decision points and any of its validating criteria, the optimal decision point was considered indeterminable.

To clarify and illustrate the analysis and reasoning used to investigate the validity of the selected IMS decision rules and to identify the apparently optimal criterion level for each objective, a description of this process is presented for two objectives. These are the objectives found in the IMS unit Fractions Level IV, and designated as objectives A-2 and A-5, on Table 2.

In Tables 4 and 5 only those decision points are presented for which there were computable chi squares, since without them no contingency coefficients could be obtained. A chi square would not be computable when any two adjacent cells of a 2 X 2 contingency table are empty. A non-computable chi square can be interpreted as a lack of any predictive power between the posttest scores dichotomized at that level and the associated validating criterion. As shown in Table 3, when there where no pupils in either a column or a row there is no relationship between the decisions made on the basis of the posttest scores and the dichotomized scores used to validate those decisions.

TABLE 3

SAMPLE 2 × 2 CONTINGENCY TABLES WHERE CHI-SQUARES ARE NOT COMPUTABLE

a)

Score on Validating Criterion		Score on Post-Test 0	Score on Post-Test 1
	1		45
	0		17

b)

Score on Validating Criterion		Score on Post-Test 0	Score on Post-Test 1
	1		
	0	20	42

1 = attained "mastery" score (advance)
0 = failed to attain "mastery" score (re-cycle)

Table 4 presents the contingency coefficients associated with Objective A-2 and the corresponding validating criteria. Note that while there is a significant relationship between decisions made on the basis of the program established mastery level and retest scores, this is not the decision point which optimizes that relationship. The level which shows the strongest significant relationships with the largest number of validating criteria is level 7. This decision point, designated on Table 4 by symbol #, is the one which the data suggest would have yielded the greatest number of correct advance/recycle decisions.

In the context of this study an instructional management decision associated with a given objective was labeled "correct" if a pupil was

Table 4

CONTINGENCY COEFFICIENTS: CONTINGENT RELATIONSHIP BETWEEN THE ATTAINMENT OF SEVERAL SUBSEQUENT PROGRAM CRITERIA AND SEVERAL LEVELS OF PERFORMANCE ON OBJECTIVE A-2

Objective A-2 Level	Re-Test	Subsequent Unit Objective B-1	B-2	B-3	B-4
9	.0701	.1641	.1173	.0197	.1289
(8)	.2878**	.1495	.1483	.1028	.1385
7#	.3672**	.2860**	.2524**	.1244	.0999
6	.2740**	.2449**	.2384*	.0822	.0487
5	.2653**	.2359*	.0910	.1415	.0804
4	.2591**	.1327	.0568	.2263*	.0171
3	.1914	.1736	.1450	.2402**	.0397
2	.2534**	.1342	.0187	.0139	.0847

N = 62

* $p < .10$

** $p < .05$

\# beside level indicates apparent optimum

() parentheses designate required program mastery level

advanced to the next unit, and then performed satisfactorily on the first set of posttests for that unit and attained the required score on the retest. A recycle decision was also labeled as "correct" if, had that pupil been advanced, he would have "failed" either one of the posttests in the next sequential unit of instruction or the retest.

Following the previously described guidelines, level 5 was identified as the apparent optimal criterion level for Objective A-5 (see Table 5). Of the contingency coefficients associated with Objective A-5, those computed using criterion level 5 are the largest and are significant for four-fifths of the validating criteria.

Utilizing these same procedures, the validity of the program mastery levels were judged. Apparent optimal criterion levels also were identified for each of the selected IMS objectives. Table 6 presents a summary of the results for all twenty of the investigated instructional management decision rules.

The results summarized in Table 6 offer evidence to support the validity of three of the twenty selected IMS instructional management decision rules. The data indicate that, for these three objectives, the program established criterion level would maximize the number of correct instructional management decisions. No change is suggested for the criterion level associated with these objectives. This is indicated by the word "none" in the appropriate column of Table 6.

For five of the selected objectives no evidence was obtained in the present investigation which would suggest that any of the possible decision points provided by the respective IMS tests could be of value in predicting subsequent pupil performance.

As shown in Table 6, the data indicated that a change in the criterion level required for the remaining twelve objectives would optimize the relationship between the decisions made on the basis of those criterion levels and the respective validating criteria. Of those twelve suggested changes, four require an increase in criterion

Table 5

CONTINGENCY COEFFICIENTS: CONTINGENT RELATIONSHIP BETWEEN THE ATTAINMENT OF SEVERAL SUBSEQUENT PROGRAM CRITERIA AND SEVERAL LEVELS OF PERFORMANCE ON OBJECTIVE A-5

Objective A-5 Level	Re-Test	Subsequent Unit Objective B-1	B-2	B-3	B-4
7	.1140	.1781	.0324	.0564	.1352
(6)	.2634**	.2236*	.1764	.1244	.1631
5#	.3546**	.3279**	.2162*	.1733	.2124*
4	.1134	.0750	.0153	.0766	.0219
3	.1911	.0283	.0261	.0342	.0738

N = 62

* p < .10

** p < .05

\# beside level indicates apparent optimum

() parentheses designate required program mastery level

Table 6

NUMBER OF TEST ITEMS, PROGRAM ESTABLISHED CRITERION LEVEL, SUGGESTED OPTIMAL CRITERION LEVEL, AND NUMBER OF PUPILS AFFECTED BY THE SUGGESTED CHANGE IN CRITERION LEVEL FOR EACH OF THE SELECTED OBJECTIVES

Objective Number	Number of Test Items	Program Estab. Criterion Level	Suggested Optimal Criterion Level	Suggested Criterion Level Change	Additional Pupils for Whom Apparently Correct Decisions Would Have Been Made	
					No.	%
A-1	8	7	I	I	--	--
A-2	9	8	7	-1	10	16.13
A-3	8	7	8	+1	10	16.13
A-4	11	9	7	-2	7	11.29
A-5	7	6	5	-1	5	8.06
A-6	3	2	I	I	--	--
A-7	6	5	4	-1	3	6.12
A-8	4	3	I	I	--	--
A-9	4	3	2	-1	7	14.29
A-10	6	5	2	-3	23	46.93
A-11	6	5	5	None	0	0
A-12	4	3	I	-2	8	25.80
A-13	7	6	5	-1	2	6.45
A-14	13	9	10	+1	4	12.90
A-15	6	5	5	None	0	0
A-16	6	5	I	I	--	--
A-17	23	20	23	+3	7	26.92
A-18	8	7	I	I	--	--
A-19	8	7	7	None	0	0
A-20	6	5	6	+1	5	19.23

I = Indeterminate

performance while the other eight indicate a lowering of the required performance levels.

Table 6 shows the number and percentage of pupils who would have been affected had the suggested criterion levels been used rather than the program established criterion level. The figures in the percentage column of Table 6 represent, for each objective, the ratio of the number of subjects affected by the suggested change in criterion level over the total number of subjects for whom data related to that objective were collected. As shown in Table 2, this last number varies depending on the objective. These figures provide an indication of the degree of instructional efficiency to be gained by using the suggested optimal criterion points. These data can be used to estimate the cost/benefit of empirically derived criterion levels.

Cost/Benefits

In investigating the value of empirical procedures for deriving criterion levels, it is necessary to examine the cost/benefits which could be expected from such procedures. This entails an estimate of the instructional efficiency which would be gained from using empirically derived criterion levels. To accomplish this some reasonable approximation is needed of: (a) the cost of each incorrect instructional decision, (b) the expected percentage of increase in the correct instructional decisions using empirically based criterion levels, and (c) the number of instructional management decisions made per year per child in the operation of the program.

The cost of each incorrect instructional decision can be estimated by multiplying the cost of each instructional period by the number of

periods consumed by each decision. The estimated cost of incorrect instructional decisions is computed here on the basis of data related to the operation of the Karlsruhe American Elementary School. The best indication of the per pupil cost of education in that school for one year is reflected in the yearly tuition fee. A tuition of $875 is charged by the Department of the Army to nongovernmental personnel wishing to enroll their dependents. As with most elementary schools, there are approximately six instructional periods per day for each of the 180 days of school, or a total of 1080 instructional periods. Based on observations and anecdotal comments there usually is about a three day period between the time a pupil is recycled and the time he attempts that post-test again. Using this information the cost of each incorrect instructional decision is estimated as:

$875 Cost/Child/ Year	÷	1080 No. of Instructional Periods/Year	=	$ .81 Cost/Child/ Instructional Period
$ .81 Cost/Child/ Instructional Period	X	3 Instructional Periods Involved in Each Decision	=	$2.43 Cost/Incorrect Instructional Decision

The results reported in section one of this chapter indicate that for twelve of the selected objectives the data suggests an optimal criterion level which is different from that required in the IMS program. Table 6 shows the number of subjects who would have been affected had the suggested optimal criterion levels been utilized instead of the program established mastery levels. A totaling of this column on Table 6

also shows that altogether 91 advance/recycle instructional management decisions would have been affected. This represents a little more than 10% of the total number of instructional decisions considered in this study. Therefore, on the basis of the data gathered in this investigation, it is estimated that, using empirically derived performance standards, the accuracy of approximately 10% of the instructional management decisions would be improved.

According to Frary (1971) the average pupil can be expected to complete about thirteen IMS units during a single school year. With approximately five objectives per unit, it is estimated that the average pupil will encounter sixty-five advance/recycle decision points during a given school year. If the use of empirically derived criterion levels would increase the accuracy by 10%, or 6.5 of these decisions, it would mean the more efficient use of about $15.80 worth of instructional time per child per year.

This dollar value is based upon an instructional period cost of $0.81, which in turn was derived from an annual per pupil expenditure of $875 for 1080 instructional periods, three of which were encumbered by each instructional decision. Variations in annual per pupil expenditures would directly affect the estimated cost per child; but variations in the length or number of instructional periods probably would not, since one is dependent on the other.

DISCUSSION

In the IMS program, as in most other individualized instructional programs, the attainment of the required criterion or mastery score on

each objective has been used to decide whether or ...il was ready to advance through the instructional sequence. Thus, the mastery levels associated with each objective functioned as decision rules used to predict subsequent pupil performance.

There were two primary purposes of the present study. One was to investigate the validity, as decision rules for predicting subsequent performance, of the mastery levels associated with a selected set of IMS instructional objectives. The other purpose was to identify, on the basis of the empirical data gathered, that criterion level for each objective which could be expected to optimize the number of correct instructional decisions.

The procedures used in this investigation appeared to be fairly successful in identifying the apparently optimal criterion level for the selected objectives. However, only three, or 15%, of the optimal criterion levels identified by the data were the same as those criterion levels intuitively established by the program. Thus, the procedures used by the instructional designers were not very accurate in identifying optimal criterion levels. Apparently, the use of empirical data can make a substantial improvement in the process of establishing criterion levels.

Many of the currently available individualized instructional programs require the same criterion or mastery level, often defined as 80% proficiency, over all objectives. The results of this investigation did not lend support to that practice.

In this investigation the suggested optimal criterion levels were, in some cases, identified on the basis of significant relationships with

only one or two of the validating criteria. There are three possible explanations why significant relationships were not obtained more frequently. First, the absence of any significant relationship between a given objective posttest and certain of the respective validating criteria might simply be reflecting the lack of any true interdependence or relationship between them. For example, there might be no relationship between a person's ability to perform on a given objective in Unit A and his subsequent proficiency on one of the specific objectives in Unit B. It is not, however, reasonable to suspect this as the cause of the often low and sometimes non-significant relationships between an objective posttest and the retest of that objective. Such low test - retest contingencies are explained by the next two factors which acted to reduce the obtained contingency coefficients.

The second factor which reduced the number of significant contingency coefficients between posttest scores and validating criteria grows from the fact that this study was conducted under regular classroom operating conditions. Although teachers were asked to make certain that after the subjects completed any Unit A posttest they did not do any further work in any IMS material for that unit, the subjects were exposed to the ongoing instructional program presented in their respective classrooms. In addition, a few teachers reported that, because it had become part of their regular operating procedures, they or their aides might have gone over missed test items with some subjects before advancing them to the next unit. The occurrence of this additional instruction could be viewed as a methodological weakness which might have been reduced had the study been conducted under more laboratory

oriented conditions. However, the presentation of this instruction is typical of the conditions that could be expected from any real classroom situation where the program had been implemented. Thus, it can also be argued that the presence of this additional classroom instruction does not detract from, but rather adds to the representativeness of the data collected from this investigation.

The third factor which served to reduce the number of significant contingency coefficients obtained was the IMS tests themselves. The study was conducted within the constraints imposed by the length, quality and thus reliability of the IMS posttests. In relation to this, it should also be noted that the selection of the optimal criterion level for any given objective was restricted to those levels made possible by the length of the corresponding IMS posttest. It is quite possible that tests could be designed which would not only have content validity, but would also have greater reliability and thus would be more sensitive to whatever relationships did exist between proficiency on the selected objectives and the corresponding validating criteria.

One of the main rationales for instruction designed around the idea of mastery learning is that at least 90% of today's learners are capable of attaining the desired goals of an instructional program if only they are given sufficient time (Bloom, 1968). Accordingly, one purported purpose of mastery learning instruction is to reduce the proportion of failure found in conventional instruction. It has been suggested that the elimination of this failure, especially in the elementary grades, would have a substantial positive effect on the self-concept and subsequent academic performance of a large segment of our school population

(Bloom, 1971). Specifically, this involves those pupils who have experienced repeated and continued failure under traditional instructional practices and grading procedures. Unfortunately, the data gathered in this investigation suggest that, although no longer graded on the basis of competition with other pupils, some pupils are still experiencing failure in mastery learning situations. In the IMS program this failure takes the form of being recycled. That recycling was interpreted as failure and had similar detrimental attitudinal and motivational effect was evidenced by anecdotal comments of teachers and pupils. These negative side-effects of recycling have also been found in other mastery learning instructional situations (Block, 1970; Lawler, 1971).

These findings emphasize the importance of identifying performance criteria which will minimize any unnecessary recycling of pupils through segments of the instructional program. In addition, if all pupils are to attain "mastery" of the required objectives in a given instructional program without experiencing the failure associated with recycling, some adaptations need to be made in that program. One possibility would be to accommodate individual pupil differences in ability through the quality or quantity of instruction provided. The quality of instruction, however, is very difficult to regulate. The most common current practice in individualized instruction is to vary the quantity of instruction largely by adjusting the amount of time instruction is presented. To minimize failure experiences these differences in the length of exposure time need to be made before the pupil takes the posttest, rather than on the basis of how many times recycling occurs. Another alternative is

to present instruction in a manner so that the pupil is unaware that recycling or remedial instruction has occurred.

The estimated increase in instructional efficiency reported in the cost/benefits section indicates that there is a substantial instructional gain to be expected from the utilization of empirically derived criterion levels in making instructional management decisions. Whether this gain is sufficient to justify the cost involved in obtaining these criterion levels, however, depends upon the conditions under which the empirical data are gathered. For the typical school or district the required data gathering and analytical procedures would be far too costly, involving resources not generally available to them.

For the identification of optimal criterion levels to be feasible and cost efficient there are at least three required conditions. These constraints can be met best by the instructional designers during a formative evaluation field test of the program. First, the process of identifying optimal criterion levels needs to be directed by someone knowledgeable in educational measurement and evaluation. Such individuals are typically found in universities and educational laboratories where instructional programs are being developed.

The second requirement is the use of a computer for instructional management and data gathering purposes. The most efficient means of gathering the large quantity of data needed for any extensive formative evaluation is through the use of computer managed instruction (CMI). While the use of CMI may or may not be required in the ordinary

operation of the instructional program, its implementation for data gathering purposes during a formative evaulation field test is highly desirable. If a computer is used during this time, it would be relatively easy to adjust the programming so that the data needed for identifying optimal criterion levels could be gathered, also.

The third requirement for the feasible and cost efficient identification of optimal criterion levels is that the cost of these procedures be amortized over a large number of users. Again, it would seem the best way to accomplish this is for the initial cost to be incurred by the instructional designers and then spread over all program purchasers.

In conclusion, the methodology and results of the present investigation suggests the need for further exploration in three additional separate, but related, areas. The first of these recommended areas of research relates to the feasibility of conducting the type of investigation reported here in a formative evaluation setting using CMI. In such a situation it might also be possible to investigate what test lengths and item characteristics function most efficiently and yield the greatest predictive as well as content validity.

The second area of research involves determining whether the optimal cirterion level for a given objective is affected by the type of learning required in that objective. In other words, before a pupil can be expected to successfully advance through a hierarchically structured instructional program, does he need to demonstrate a different degree of proficiency on an objective requiring the attainment of verbal information than on an objective requiring the attainment of verbal information than on an objective requiring the use of some intellectual skill? The most

appropriate context for answering this question might be a science curriculum, due to the large number of both types of learning required in that area.

Third, educators, and especially instructional designers, need to give more consideration to a number of important instructional questions: What is mastery? Why is mastery, as defined in the program, required for a particular objective? By what standards can instructional designers judge whether the attainment of "mastery" on a given objective served the desired purpose? To a very real degree, this involves identifying criteria which can be used to judge, at least partially, the value as well as the effectiveness of requiring mastery on particular objectives.

SUMMARY

Current discussions about "mastery" learning and criterion referenced instruction reflect great diversity in the way criterion (or "mastery") levels have been defined and determined. Criterion levels within a program are viewed as decision rules used to judge when a pupil is ready to advance to the next step in an instructional sequence.

The purposes of this study were to investigate the validity, as instructional management decision rules, of the preset en-route criterion levels (ECL) associated with a selected group of instructional objectives in an individualized mathematics program and to test procedures for empirically determining optimal ECLs. The validation was conducted in terms of delayed retest scores and in terms of subsequent progress through the instructional continuum. Subsequent progress was operationally defined as attaining the required performance level on the objectives

found in the next sequential unit of instruction. The empirical data gathered were used to suggest an optimal criterion level for each of the selected objectives.

This investigation was conducted in the context of the Individualized Mathematics System (IMS) as implemented in the American Elementary School in Karlsruhe, Germany. Subjects were selected from those pupils working on the selected subset of IMS objectives during the time of the study.

To investigate the validity of the decision rules, expressed as mastery levels associated with the objectives in a particular unit, the subjects who took the posttest for that unit were advanced to the next sequential unit without regard to their posttest scores on the first unit. A retest of each selected objective was administered as soon after the subjects completed the next sequential unit as possible. The data collected consisted of posttest scores for each of the selected objectives, delayed retest scores on these objectives, and posttest scores for each objective in the next sequential instructional unit.

The decision point selected as optimal for each objective was the one which yielded the largest significant contingency coefficients with the greatest number of validating criteria. The validity of the a priori mastery level established for each objective was judged on the basis of whether it was the one selected as optimal for that objective.

The procedures used in this investigation appeared to be successful in identifying optimal criterion levels for the selected objectives. In terms of validating the program mastery levels, however, only 15% of the optimal criterion levels identified by the data were the same as those criterion levels originally established by IMS. Thus, the procedures used

by the instructional designers were not very accurate in identifying optimal criterion levels. Apparently, the use of empirical data can make a substantial improvement in the process of establishing en-route criterion levels.

The increase in instructional efficiency estimated in the cost/benefits section of this report indicates that there is a substantial instructional gain to be expected from the utilization of empirically derived optimal criterion levels. For this gain to jusify the cost involved in obtaining these criterion levels, however, it is suggested that the data be collected by the instructional designers during the formative evaluation field test of the program, using automated data processing and multiple-matrix sampling techniques.

REFERENCES

Airasian, Peter W. The role of evaluation in mastery learning. In James H. Block (Ed.), Mastery learning: theory and practice. New York: Holt, Rinehart and Winston, Inc., 1971.

Biehler, R. F. A first attempt at a "learning for mastery" approach. Educational Psychologist, 1970, 7 (3), 7-9.

Block, James H. The effects of various levels of performance on selected cognitive, affective, and time variables. Unpublished doctoral dissertation, University of Chicago, 1970.

_____. Operating procedures for mastery learning. In James H. Block (Ed.), Mastery learning: theory and practice. New York: Holt, Rinehart and Winston, Inc., 1971.

Bloom, B. S. Learning for mastery. Evaluation Comment, 1968, 1 (2).

_____. Affective consequences of school achievement. In James H. Block (Ed.), Mastery learning: theory and practice. New York: Holt, Rinehart and Winston, Inc., 1971.

Bloom, B. S.; Hastings, J. T.; & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill Book Company, 1971.

Bolvin, J. O.; Lindvall, C. M.; & Scanlon, R. G. Individually prescribed instructional manual. Pittsburgh: University of Pittsburgh, 1967.

Cooley, W. H. and Glaser, R. An information and management system for individually prescribed instruction. In R. C. Atkinson and H. A. Wilson (Eds.), Computer-assisted instruction: a book of readings. New York: Academic Press, 1969.

Coulson, J. E., and Cogswell, J. F. Effects of individualized instruction on testing. Journal of Educational Measurement, 1965, 2, 59-64.

Cronbach, L. J. Validation of educational measures. In R. L. Thorndike (Ed.), Educational Measurement. American Council on Education, 1970.

Flanagan, J. C. Project PLAN. In Technology and innovation in education. Prepared by the Aerospace Education Foundation. New York: Praeger Publishers, 1968.

Frary, R. B. Formative evaluation of the individualized mathematics system (IMS). Durham, North Carolina, National Laboratory for Higher Education, 1971.

Gagné, R. M. The analysis of instructional objectives for the design of instruction. In Robert Glaser (Ed.), Teaching machines and programmed learning. Washington, D. C.: National Education Association, 1965.

_____. Learning hierarchies. Educational Psychologist, 1968, 6 (1), 1-6.

Gagné, R. M.; Mayor, J. R.; Garstens, H. L; & Paradise, N. E. Factors in acquiring knowledge of a mathematical task. Psychological Monograph, 1962, 76, NO. 7 (Whole No. 526).

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

_____. Adapting the elementary school curriculum to individual performance. In Proceedings of thd 1967 invitational conference on testing problems. Princeton: Educational Testing Service, p. 3-36, 1968.

Hackett, M. G. Success in the classroom--an approach to instruction. New York: Holt, Rinehart & Winston, 1971.

Ironside, E. M. (Ed.) IMS user guides. Durham, N. C.: National Laboratory for Higher Education, 1971.

Kersh, Mildred E. A strategy for mastery learning in fifth-grade arithmetic. Unpublished doctoral dissertation, University of Chicago, 1970.

Kriewall, T. E. Application of information theory and acceptance sampling principles to the management of mathematics instruction. Technical report No. 103, October 1969, Wisconsin Research and Development Center, Madison.

Lawler, R. M. An investigation of selected instructional strategies in an undergraduate computer-managed instruction course. Unpublished doctoral dissertation, The Florida State University, 1971.

Mager, R. F. Preparing instructional objectives. Palo Alto: Fearon Publishers, 1962.

Mayo, S. T.; Hunt, R. C.; and Tremmel, F. A mastery approach to the evaluation of learning statistics. Paper presented at annual meeting of National Council on Measurement in Education, Chicago, Illinois, 1968.

Merrill, M. D. Correction and review on successive parts in learning a hierarchical task. Journal of Educational Psychology, 1965, 56, 225-234.

_____. Measuring learning outcomes. In M. D. Merrill (Ed.), Instructional design: readings. Englewood Cliffs, N. J.: Prentice-Hall, 1971.

Popham, W. J., and Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Postlewait, S. N.; Novak, J. D.; & Murray, H. An integrated experience approach to learning with emphasis on independent study. Minneapolis: Burgess Publishing Company, 1970.