DOCUMENT RESUME

ED 120 250	95	TM 005 212

AUTHOR	Nitko, Anthony J.
TITLE	Problems in the Development of Criterion-Referenced Tests: The IPI Pittsburgh Experience.
INSTITUTION	Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	74
NOTE	27p.; Footnotes may not reproduce clearly due to small type
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS	Achievement; Behavioral Objectives; Content Analysis; *Criterion Referenced Tests; Instructional Systems; Task Performance; *Test Construction; *Testing Problems

ABSTRACT

These four characteristics inherent in criterion-referenced tests form the central theme of this paper: (1) The classes of behaviors that define different achievement levels are specified as clearly as possible before the test is constructed; (2) Each behavior class is defined by a set of test situations (that is, test tasks) in which the behaviors can be displayed in terms of all their important nuances; (3) Given that the classes of behavior have been specified and that the test situations have been defined a representative sampling plan is designed and used to select the test tasks that will appear on any form of the test; and (4) The obtained score must be capable of expressing objectively and meaningfully the individual's performance characteristics in these classes of behavior. The focus of this paper is on the development of criterion-referenced tests having these four properties and some associated technical problems that are encountered. Solutions for these technical problems are not readily available nor immediately generalizable to all curricular areas for which criterion-referenced tests might be desired. Attempts are made, therefore, to specify procedures that will be useful to the practical developer until the technical problems are solved. (RC)

LEARNING RESEARCH AND DEVELOPMENT CENTER

1974/22

PROBLEMS IN THE DEVELOPMENT OF CRITERION-REFERENCED TESTS: THE IPI PITTSBURGH EXPERIENCE

ANTHONY J. NITKO

University of Pittsburgh

PROBLEMS IN THE DEVELOPMENT OF CRITERION-REFERENCED TESTS: THE IPI PITTSBURGH EXPERIENCE

Anthony J. Nitko

Learning Research and Development Center

University of Pittsburgh

Reprinted with permission from C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.), Problems in Criterion-Referenced Measurement. (CSE Monograph Series in Evaluation 3.) Los Angeles: University of California, Center for the Study of Evaluation, 1974. Pp. 59-82.

Preparation of this paper was supported by the Learning Research and Development Center, which is supported in part by funds from the National Institute of Education (NIE), United States Department of Health, Education and Welfare. The opinions expressed do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred.

PROBLEMS IN THE DEVELOPMENT OF CRITERION-REFERENCED TESTS: THE IPI PITTSBURGH EXPERIENCE

Anthony J. Nitko
Learning Research and Development Center
University of Pittsburgh

Criterion-referenced testing has come about on a new wave of psychology—a psychology expressing an increasing concern for instruction and the instructional process. Such an instructional psychology postulates a theory of instruction that is prescriptive with respect to the instructional procedure itself. A learning theory, on the other hand, is descriptive and after the fact, specifying the conditions under which the learning occurred (Burner, 1966).

> In theories of instructional psychology primary focus is on . . . (a) a description of the state of knowledge to be achieved; (b) description of the initial state with which one (i.e., the learner) begins; (c) actions which can be taken, or conditions that can be implemented to transform the initial state; (d) assessmeat of the transforaiation of the state that results from each action; and (e) evaluation of the statement of the terminal state desired (Glaser & Resnick, 1972, p. 208).

Glaser's motive for applying criterion-referenced testing to educational achievement measurement (Glaser, 1963) stemmed from a concern about the kind of achievement information needed to make instructional decisions from the above kind of instructional psychology. Some instructional decisions concern individuals and may relate, for example, to the kind of competence an individual needs in order for him to be successful in the next course of a sequence. Other decisions center around the adequacy of the instructional procedure itself. Tests that provide achievement information about an individual only in terms of how the individual compared with other members of the group tested, or which provide only sketchy information about the degree of competence the individual possesses with respect to some desired educational outcome, are not sufficient to make the kinds of decisions necessary for effective instructional design and guidance.

Glaser's (1963) application combined both the notion of a desired model

Preparation of this paper was supported by the Learning Research and Development Center, which is supported in part by funds from the National Institute of Education (NIE), U.S. Department of Health, Education and Welfare. The opinions expressed do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred. Grateful acknowledgement is made to Drs. Cooley, Cox, Glaser, Hsu, and Resnick for their helpful comments on the draft manuscript.

or a minimum goal we would like an individual to attain (Flanagan, 1951) and the notion of a standard domain of content (Ebel, 1962). He called for the specification of the type of *behavior* the individual is required to demonstrate with respect to the content. This distinction between behavior or performance and content is at the heart of criterion-referenced testing. "The standard [or criterion] against which a student's performance is compared . . . is the behavior which defines each point along the achievement continuum (Glaser, 1963, p.519)." *A criterion-referenced test, then, is one that is deliberately constructed to give scores that tell what kinds of behavior individuals with those scores can demonstrate* (Glaser & Nitko, 1971).

Note that this definition does not imply a predetermined, fixed cutting-score (cf. e.g., Livingston, 1972); it does not imply simply writing a set of behavioral objectives and keying a set of items to those objectives; and it does not imply the use of only open-ended production items (cf. Harris & Stewart, 1971). The definition, instead, implies that there are four characteristics inherent in criterion-referenced tests:

> The classes of behaviors that define different achievement levels are specified as clearly as possible before the test is constructed.
>
> Each behavior class is defined by a set of test situations (that is, test tasks) in which the behaviors can be displayed in terms of all their important nuances.
>
> Given that the classes of behavior have been specified and that the test situations have been defined, a representative sampling plan is designed and used to select the test tasks that will appear on any form of the test.
>
> The obtained score must be capable of expressing objectively and meaningfully the individual's performance characteristics in these classes of behavior (Nitko, 1970).

These four characteristics form the central theme of this paper. The focus is on the development of criterion-referenced tests having these properties and some associated technical problems that are encountered. Solutions for these technical problems are not readily available nor immediately generalizable to all curricular areas for which criterion-referenced tests might be desired. Attempts are made, therefore, to specify procedures that will be useful to the practical developer until the technical problems are solved.

The characteristics outlined above appear to form a logical developmental sequence. This sequence is seldom followed in practice. In fact, a great deal of criterion-referenced test development is still in the intuitive or artistic state. More often than not the procedure is iterative. For example, attempts to specify classes of behavior may begin by first specifying varieties of test items. These items might be subjected to behavioral analysis and behavioral class descriptions are then induced. This may lead to further specification of items or redefinition of behavior classes.

Permeating all of the discussion that follows is the notion of a theory of performance (Miller, 1962; Hively, 1970) or an analysis of the psychological processes underlying task performance. This type of process analysis is used to structure the classes of behavior defining various levels of achievement and in interpreting specific item performance as representing the class of behavior defined.

DOMAIN DEFINITION

Of the four characteristics of criterion-referenced testing[1] outlined earlier, specifying classes of behavior that define different levels of achievement is the most difficult to achieve. The failure to adequately specify this domain has led to recent criticisms of criterion-referenced testing (e.g., Ebel. 1970; Stanley & Hopkins, 1972). Since these criticisms hark back to the inadequacy of the old percentage grading system, perhaps the demise of that system was also due to the domain specification failure.

A complete exposition on domain specification is beyond the scope of this paper. It is useful, however, to sketch out some of the dimensions of the problem so that the practical developer of criterion-referenced tests may take them into consideration. These dimensions include establishing various levels of achievement, the relationship between ultimate and proximate achievement levels, the nature of the domain specification, and the derivation of domain descriptions.

Levels of Achievement

Performance or achievement criteria can be established at any convenient point in the instructional process. For example, the classes of behavior defining various levels of competence can be specified at the termination of a course, at the termination of a unit of instruction (i.e., smaller within-course segments of instruction), or at any other point during the course of instruction. The definition of these behavior domains will be guided by the nature of the instructional system and the purpose for which the information will be used, e.g., certification of attainment, within curriculum placement, or diagnosis of deficiencies (cf. Glaser & Nitko, 1971).

At the termination of instruction broad domains of performance are definable. The definition and analysis of these domains occur at several levels ranging from the definition of the desired outcomes of the entire

[1] While it may be useful for some to avoid the term criterion-referenced testing and focus on criterion-referenced score interpretation (e.g., Simon, 1969; Davis, 1970), it seems more useful to refer to "tests" in the context of this paper. In order to have criterion-referenced score interpretation, scores need to be referenced back to the behavior domain. Hence, focus in development should be primarily on the domain of behavior and the derivation of test tasks to elicit that behavior, rather than short-cutting these and focusing mainly on the scores (cf., Jackson, 1971).

educational enterprise, at one extreme, to the specification of the desired outcomes at the termination of a particular subject-matter course, at the other extreme. The former is likely to yield many domain definitions, be divergent, and require many tests in order to assess pupil outcomes. The latter leads to fewer domains, is more convergent in terms of outcome categories, and may result in fewer tests.

Ultimate and Proximate Behavior

Defining levels of achievement at various points in instruction raises the issue of what kinds of behavior are important enough to be included in a domain specification. While this is an old area and subject to considerable debate and discussion, it is not yet resolved. The importance of the distinction between proximate and ultimate objectives of instruction for educational test developers was articulated several years ago by Lindquist (1951).

Educational practice generally assumes that the knowledge and capabilities with which the learner leaves the classroom are related to the educational goals envisioned by society. This assumption implies that the long-range goals the learners are to attain in the future are known and that the behaviors with which the learners leave a particular course actually contribute to the attainment of these goals. What is closer to reality, however, is that the long-term relationship between what the student is taught and the way he is eventually required to behave in society is not very clear (Glaser & Nitko, 1971).

In contrast to ultimate goals, proximate goals define the domains of performance that a learner displays at the end of a particular instructional situation (e.g., course or grade level). It should be noted that proximate objectives are not defined as the materials of instruction nor as the particular sets of test items that have been used in the instructional situation. For example, at the end of a course in spelling one might reasonably expect a student to be able to spell certain classes of words from dictation. During the course, certain of these words might have been used as examples or as practice exercises. The instructor would be interested in the student's performance with respect to the class or domain of words as a proximate objective of instruction and not the particular words used in instruction. Thus, to assess a student's performance with respect to a domain, *one may need to consider the transfer relationship between the items in the domain and the preceding instruction.*

General Nature of Domain Specification

The specification of the domain of instructionally relevant achievement behaviors can profit much from the suggestions for "universe specification" advocated by Cronbach (1971). As Cronbach has pointed out, too often attention is paid only to the selection of subject-matter topics. The nature of the stimulus and the description of the response are ignored. Proper domain specification requires that both stimulus and response descriptions be included. Thus,

> A proper response specification deals with the result a person is asked to produce, not the process(es) by which he succeeds or fails. 'Reads printed words aloud' is a description of an observable response; it says nothing about whether the reader is to look and say or to sound the word out. A person who insisted on separating these two response processes would have to devise a new task specification, perhaps requiring the reading of nonsense constructions that no subject has seen before. If a category of the form say, 'ability to evaluate arguments' is to mean anything as a task specification, the designation must be fleshed out to describe something about the stimulus, the accompanying injunction to the subject, and the aspect of the behavior to which the scorer is directed to attend (Cronbach, 1971, p. 454).

In this sense, use of the *Taxonomy of Educational Objectives* (Bloom, 1956) is insufficient for domain specification since the categories described therein are inferred psychological processes. However, to adequately specify the dimensions of the performances to be included in the domain, one may need to invoke a theory of performance (Hively, 1970; Miller, 1962) to decide which stimulus and response characteristics are relevant for domain description. This point will arise again when deriving tasks from the behavior description is discussed.

Derivation of Domain Description

While in practice the generation of performance domains is often ultimately tied to the actual specification of the tasks (stimuli) themselves, this derivational process is discussed separately here. It should be noted, however, that the state of the technology for determining the content and attributes of *what* is learned is not well developed, particularly where behavioral characteristics of complex school-like performances is concerned (Glaser & Resnick, 1972).

One practical method for deriving domain descriptions for smaller classes of behavior, such as a domain of behavior relevant for a unit[2] of instruction, is the procedure stemming out of Gagné's work on learning hierarchies (e.g., Gagné & Paradise, 1961). [A modification of this procedure, which seems to give more replicable results, has been provided by Resnick (Resnick, Wang, & Kaplan, 1970).] The analysis of learning hierarchies begins with any desired instructional objective, behaviorally stated, and asks in effect: To perform this behavior what prerequisite or component behaviors must the learner be able to perform? For each behavior so identified, the same question is asked, thus generating an ordered hierarchy

[2] The analysis of learning hierarchies need not be restricted to units of instruction, of course. It may be possible to apply the procedure to broad curricular areas.

of behaviors based on testable prerequisites. The analysis can begin at any level and always specifies what comes earlier in the curriculum. It should be noted that as it is used here, hierarchy analysis is a tool for domain definition. Whether all students' learning should progress through the hierarchy in the same way is an empirical question for instructional psychology.

As a result of this type of analysis and domain specification, the test developer is provided with the essential information about what behaviors are to be observed and tested in order to determine the status of the learner with respect to the achievement continuum. Thus a hierarchical analysis provides a good map on which the attainment, in performance terms, of an individual student may be located. The uses of such hierarchies in designing a testing program for a particular instructional system are described elsewhere (Glaser & Nitko, 1971).

A serious question that can be raised is how much of education can be analyzed into hierarchical structures. The answer to the question is very much an open, experimental matter. Three things should be noted, however (Glaser & Nitko, 1971). First, the development of hierarchies for complex behaviors may lead to several such structures, each of which is "valid" with different kinds of learners, but none of which, taken alone, is valid for all learners. Second, the analysis of behaviors into components and prerequisites leads to structures that stand as hypotheses open to empirical verification. Third, in actual instructional practice there is always a functional sequence wherein the instructor has at least an intuitive hierarchy through which he proceeds.

Another point to remember is that criterion-referenced interpretations are most useful when the behavior domain has an orderly progression (Cronbach, 1970). Hierarchy analysis, or a similar procedure, would seem to be a useful tool in discovering these progressions.

The use to which the test is to be put will to a large extent determine the nature of the performance to be included in the domain definition. For example, one may develop performance domains by analysis of an "expert's" behavior or by the analysis of an "amateur's" behavior (Hively, 1970). It may well be that certain elements of performance will drop out as task proficiency increases. For assessment of initial stages of learning, therefore, it may be that more components need to be included in the domain definition (and consequently on the test) than at later stages of learning. This would seem to imply a distinction between diagnosis, placement, and final (terminal) learning assessment (see Glaser & Nitko, 1971).

DEFINING CLASSES OF ITEMS

Closely associated with the definition of behavior classes related to levels of achievement is the translation of these behavioral statements into sets of test situations—test tasks or test items. Although discussed here separately, in practice these two steps are often iterative. Performance domains tend to be verbal statements and descriptions (e.g., behavioral objectives) whereas test situation descriptions tend to be more concrete in that the characteristics of the testing situation and the various type of admissible test items are mapped out and specified. Test items here refer to any carefully described ". . . stimulus conditions under which a student is expected to respond, together with the specifications for recording and scoring his response when it occurs (Hively, 1970)." Items include both performance and traditional paper-and-pencil types of items as long as these are derived from the domain definitions.

Item Forms

A useful tool for criterion-referenced tests is item forms analysis (Hively, 1966; Hively, Maxwell, Rabehl, Sension, & Lundin, 1973; Hively, Patterson, & Page, 1968; Osburn, 1968). Item forms analysis is a variation on task analysis. It is the process whereby behavioral statements are analyzed in order to derive classes of items which elicit the various aspects of the behavior class. As a result of this analysis, one or more item forms are derived for each behavior class. An item form consists of a specification of the invariant part of the class of items together with (a) an indication of which parts of the items are variable, (b) a specification of elements which can be used in the variable parts of the items, and (c) a specification of the rules by which one selects an element from the set of variable elements to derive a particular item (Hively, 1970; Hively, et al., 1973). The variant part of the item is called a *shell;* the sets of elements which can be used in the variable parts are called *replacement sets;* and the rules by which one samples from the replacement sets are called the *replacement structure* (Hively, 1970; Hively, et al., 1973).

In practice, one often cannot go directly from a verbal statement of a behavior class to an item form. The procedure usually is to first develop prototype items admissible as test tasks under the described behavior. Process and component analysis (cf. Resnick, Wang, & Kaplan, 1970) of these prototype items often leads to a modification of the original behavior specification, elimination of some of the prototype items as not implied by the behavior class, or a rewriting of the prototype items. In examining these prototype items to determine their fit to the behavioral definition one invokes a behavioral analysis and a theory of performance. This process involves more than superficial judgment and sorting. The questions that need to be answered are: (1) Does this item contain the stimulus characteristics implied by the behavioral statement? (2) Will the examinee's response to this item be indicative that he indeed has the desired response in his repertoire?

Once the set of prototype items has been delineated item forms can be induced. The pmtotype item is one member of the class of items implied by an item form. The task here is to identify the general form (format) of the items, the item shell, the variable elements, and the admissible replacement sets. Again, this implies a behavioral analysis and a theory of performance.

Item Tryout Data

As part of the procedure for defining test tasks that are consistent with domain definitions, it is necessary to establish empirical procedures for tryout of items. A major purpose of traditional item-tryout procedures is to collect data necessary to improve the test items. This is no less true when criterion-referenced test items are developed.

Tryout of items for criterion-referenced test development seeks to further refine and polish the domain of test tasks. All the ambiguities that are inherent in traditional item writing are inherent in criterion-referenced item writing. Further, since item forms are developed using behavioral analysis and performance theory, the data from item tryout are used to check on the adequacy of this initial analysis. Often this will lead to a respecification of the item form or one or more of its components —replacement sets or replacement structure (cf. Osburn, 1968).

There are those who advocate either explicitly (e.g., Stenner & Webster, 1971) or implicitly (e.g., Baker, 1971) that items designed to test a specific class of behaviors be homogeneous. Homogeneous tends to be defined in terms of item and total test score parameters such as discrimination indices and internal consistency reliability estimates. These correlation-related indices tend to be maximized when each item measures the same factor (process) (Lord, 1958). The insistence on homogeneity in this sense is too sweeping and is poor psychology. It leads to statistical techniques being used to drive the definition of performance domains. There is no logical basis for contending a priori that any domain of performance identified as instructionally relevant ought to be homogeneous (cf. Cronbach, 1971). Homogeneity should be viewed as a question for empirical experimentation and item performance theory (cf. Bormuth, 1970) and would probably vary with the target population and the class of behaviors under consideration. Heterogeneity would mean that a larger number of observations are needed before adequate generalizations about domain performance can be made.

Hierarchy Validation

If hierarchy analysis is used to develop the test domain, empirical data needs to be collected to validate this structure in terms of the items defining the various levels of the hierarchy. One should distinguish what might be called the "psychometric" hierarchy[3] from the learning hierarchy.

[3] For an example of procedures used to validate psychometric hierarchies see Wang, Resnick and Boozer (1971) and Ferguson (1970).

Classes of test tasks (items) can be ordered in hierarchical ways which may bear little relationship to the sequence in which learning should proceed. If the hierarchical ordering of the domain implies an instructional sequence, or if it represents a hypothesis about behavioral acquisition derived from instructional theory, then empirical transfer studies are required as well. Thus, criterion-referenced testing is not exempt from construct validation studies (cf. Cronbach, 1970).

Item Performance and Instruction

An important consideration in the tryout of test items in this context is the relationship between instruction and the test item domain. The tryout data is dependent on: "(1) the characteristics of the item itself, (2) the program of instruction with which it is associated, (3) the sample of the students from whom the data were collected, and (4) the conditions under which the students worked (Hively, 1966, p.7)." These are factors which influence the interpretation of tryout data and the subsequent decisions that are made concerning item and domain revision.

If the behavioral domain and subsequently derived item classes are based on some inferred process (e.g., application in the Bloom *Taxonomy*) or an inferred psychological construct (e.g., a hierarchy of prerequisite behaviors), then the content and nature of the examinees' previous learning history (i.e., instruction) need to be considered in interpreting tryout data. A similar point is made by Bormuth (1970) who calls for the development of procedures for relating the structure of the items to the structure of the instruction. For example, to adequately derive classes of test tasks measuring transfer, application, and evaluation behaviors it is necessary to eliminate from the item form those items on which the students were given practice, thus leaving those items that elicit responses not explicitly taught, but which can be deduced from instruction. Without such procedures, it is not possible to determine whether the classes of items are indeed achievement items, as opposed to general knowledge or aptitude items.

The development of items for criterion-referenced tests and the associated empirical data generated by tryout and study of these classes of items seem to call for aspects of achievement test theory that are as yet not well developed. Bormuth labels these *item-writing theory* and *item-response theory*. Item-writing theory would lead to the development of procedures for defining classes of items (item forms) and item-response theory would lead to explanations of the processes that account for responses to classes of items. The developer of criterion-referenced tests should refer to Bormuth's book for suggestions along these lines and for indications of some of the problems involved in pursuing research in these areas. It should be emphasized that theories and research in these areas are currently inadequate or completely lacking.

SELECTING ITEMS TO APPEAR ON THE TEST

Once the behavior domain and the classes of items have been specified the final stages of test development can proceed. It might be argued that the preceding discussion concerning domain definition is no more than what any test developer should do in order to maximize content validity, regardless of whether a criterion-referenced or a norm-referenced test is to be developed. While this is probably more of a fond hope than a reality, one is still inclined to agree that perhaps all test developers should take such care in developing tests. It should be noted, however, that content validity implies an indication of the sampling plan by which the particular items that appear on a particular test form are selected from the domain of all items (Cronbach, 1970).

It is assumed here that empirical data and performance theory support the definitions of achievement levels in the domain and the classes of test tasks operationalizing these behavior classes. The task is to select items to put on a form of the test in such a way that performance on that test will be a basis for an inference about the examinee's performance in the domain. It has already been mentioned that criterion-referenced test score interpretation is most meaningful when the behavior domain has an orderly progression. This implies taking advantage of the psychological structure of the subject-matter domain in selecting test items.

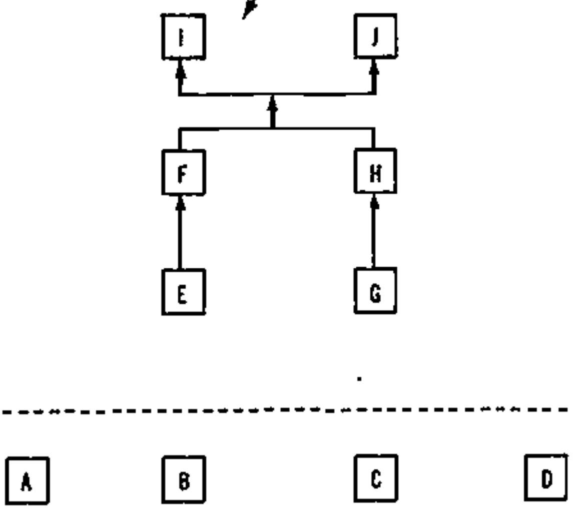
Examples of Item Selection for Curriculum Placement

If an instructional system is adaptive, it will avoid teaching the student that which he has already learned and will instead offer him new goals to learn. Information is needed to answer the question, "Where in the instructional sequence should the student begin his study?" Tests built to provide this information are specific to the content and psychological structure of the particular course of instruction with which the student is faced.

In broad areas such as an entire course or an entire curriculum area, neat hierarchies of the Gagné type covering the entire course of instruction may not exist or may become very complicated. Nevertheless, some sequencing of instructional objectives is possible. An illustration of this is shown in Figure 1 in which an elementary school mathematics curriculum has been defined in terms of approximately 350 instructional objectives. The content has been broken down into ten topics which are roughly in a prerequisite order (from top to bottom in the figure). Further, each topic has been developed over a range of complex behaviors that are also in a rough prerequisite order (from Level A through Level G in the figure). Each cell of the grid represents several instructional objectives and is called a unit of instruction. The objectives in a unit of instruction can usually be arranged in a hierarchy that leads to a few terminal goals for that unit. The inset shows (hypothetically) how a short sequence of objectives might look for one unit of instruction. Within a single unit, in general, there will be prerequisite behaviors from earlier topics and lower levels. These are labeled as behaviors A, B, C and D in the inset.

One way to place a pupil in this curriculum is to develop a two-stage

Content (Topic)	Level of Complexity						
	A	B	C	D	E	F	G
Numeration/Place Value	*	*	*	*	*	*	*
Addition/Subtraction	*	*	*	*	*	*	*
Multiplication		*	*	*	*	*	*
Division		*	*	*	*	*	*
Fractions	*	*	*	*	*	*	*
Money	*	*	*	*			
Time	*	*	*	*	*		
Systems of Measurement		*	*	*	*	*	*
Geometry		*	*	*	*	*	*
Applications		*	*	*	*	*	*



*Indicates a unit of instruction consisting of one or more instructional objectives.

Figure 1. Example of Curriculum Layout for Individually Prescribed Instruction Elementary Mathematics

MATHEMATICS PLACEMENT PROFILE

Name John Smith Date 5/20 Grade 5

School Sweetdate Teacher Mrs. Jones Room 12

Mathematics Area	Placement Level A-G							Placed at Level
	A	B	C	D	E	F	G	
Numeration/Place Value	▨	▨	▨	▨				E
Addition/Subtraction	▨	▨	▨	▨	▨			F
Multiplication	▨	▨	▨	▨				E
Division	▨	▨	▨					D
Fractions	▨	▨	▨					D
Money	▨	▨	▨	▨				--
Time	▨	▨	▨	▨	▨			--
Systems of Measurement	▨	▨	▨	▨	▨			F
Geometry	▨	▨	▨	▨				E
Applications	▨	▨	▨					D

Figure 2. Example of Placement Profile for a Hypothetical Student with Respect to the Mathematics Curriculum of Individually Prescribed Instruction

placement test (Cox & Boston, 1967). The first-stage test is broad-ranged over the curriculum. The results are used to place a student at a unit in each topic or content area. The second-stage test is narrow-ranged and tests the domain of behavior implied by a single unit. The results are used to place a student at a particular objective within a unit. The first-stage test needs to be administered only once at the beginning of a course of study. After completing instruction on the first unit of study, the student is given the second-stage test for the next sequential unit. Thus, he is placed at each successive unit in the curriculum. Figure 2 shows a completed first-stage placement profile for a hypothetical student. Figure 3 shows what a completed second-stage placement profile might look like.

The broad-range test is actually a battery of tests consisting of one test for each topic. Each subject would predict for each topic the last unit in the sequence from A to G in which the student would be successful. Traditional item-selection procedures that seek to maximize predictive validity would seem appropriate for this type of broad-range test. If the behaviors defined within a unit are hierarchical, then one could select

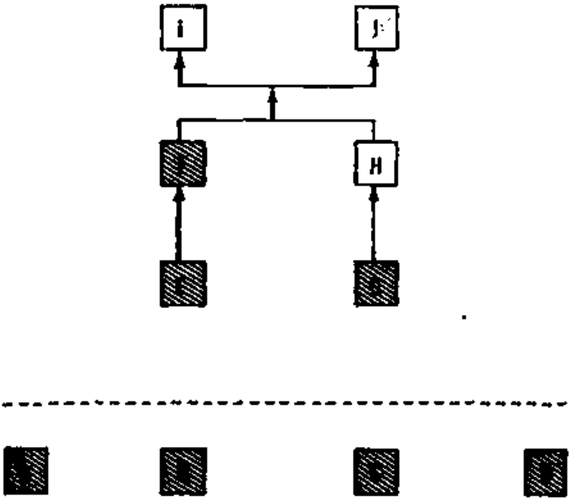


Figure 3. Placement Profile for a Hypothetical Student (Shaded boxes mean that the student has sufficient mastery of these instructional goals to proceed with a new instructional goal.)

items from the domains that define the terminal objectives for that unit, and depend on the prerequisite nature of the hierarchy to subsume the other behaviors in the unit. If a within-unit hierarchy does not exist, then selecting items from the domains of all the within-unit behaviors would seem to be required. Care should be taken, however, in using correlational indices for this type of prediction; it is the absolute level of attainment of unit skills that is of prime importance.

The second-stage type of unit test serves as another example of how items might be selected by taking advantage of the psychological structure of the subject-matter content. If the unit behaviors are hierarchical and domains of items are defined for each node in the hierarchy, then a branched test can be used to obtain a pupil's profile with respect to this hierarchy. Thus, if an examinee was successful on items testing one objective in the hierarchy, this would indicate that items from earlier objectives in the hierarchy would be passed as well.[4] Procedures for branched testing initially proposed by Ferguson (1970) and further elaborated by Hsu (Ferguson & Hsu, 1971; Hsu & Carlson, 1972) have been successfully used in an elementary mathematics curriculum when coupled with item forms and a computer.

[4] Such elaborate procedures would have to be balanced out against efficiency criteria. For example, in small hierarchies consisting of a few nodes a tailored test would be more elaborate than necessary. A student might be placed more quickly and efficiently by simply testing all nodes.

Figure 4 is a schematic illustration of terminal and prerequisite instructional objectives for an addition-subtraction unit from the elementary arithmetic curriculum of the Individually Prescribed Instruction Project (Lindvall & Bolvin, 1967). Each box represents one objective. The objectives are arranged in a branched hierarchy. Objectives 6, 17, and 18 are terminal objectives for the unit; the remaining objectives are prerequisites. Each of these prerequisites and terminal objectives is defined by one or

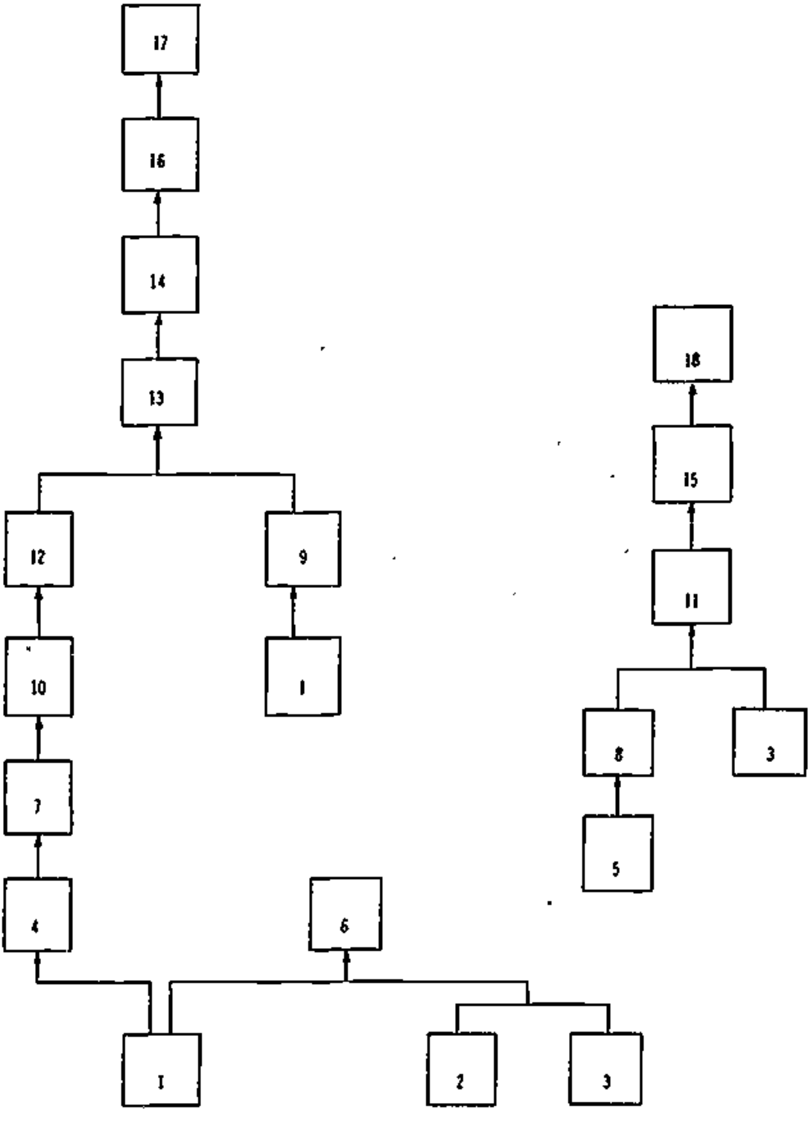
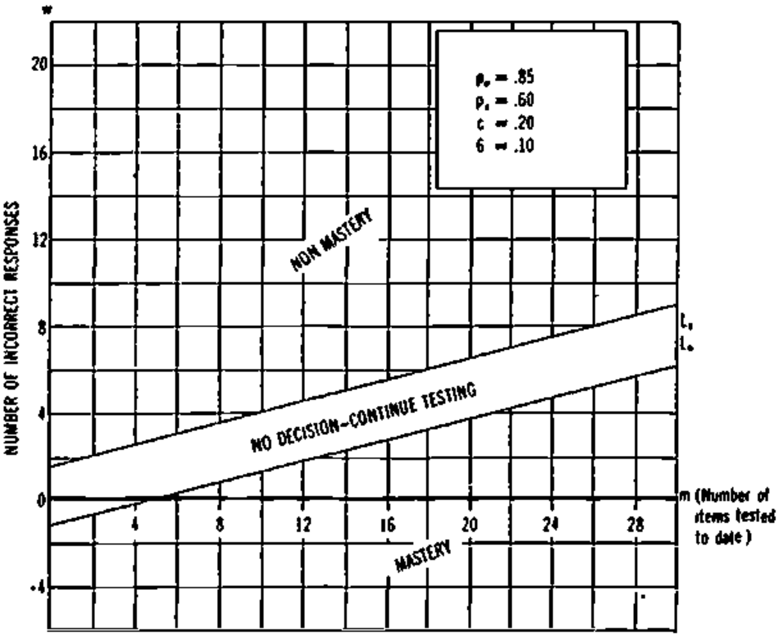


Figure 4. An Example of a Hierarchy of Skills in an IPI Mathematics Unit

more item forms which are then programmed for use on the computer. The testing is done on an individual basis at a computer terminal.

The object of the testing scheme is to locate a pupil at one of these objectives or "boxes" as quickly as possible and in such a way that he demonstrates mastery of objectives below his location and non-mastery of objectives above his location. The decisions for which the testing procedure must provide information are (1) what objectives should be tested and (2) whether the pupil has mastery or non-mastery[5] of the objectives that are tested. A decision needs to be made about every objective, but the trick is to make these decisions without testing every objective, and to minimize the testing for those objectives that are tested.

On this basis, a set of decision rules is devised that combines the capabil-



H_0: p = .85 (Student has sufficient mastery, omit instruction)

H_1: p = .60 (Student does not have sufficient mastery, give instruction)

Figure 5. Graph Illustrating Sequential Probability Ratio Test for Determining Whether a Student Does or Does Not Need Instruction on an Objective (Modified from Ferguson, 1970)

[5] By mastery it is meant that ". . . an examinee makes a sufficient number of correct responses on the sample of test items presented to him in order to support the generalization (from this sample to the domain or universe of items implied by an instructional objective) that he has attained the desired, pre-specified degree of proficiency with respect to the domain (Glaser & Nitko, 1970, p.641)."

ities of the computer with statistical logic and subject-matter logic. This allows "on-line" decisions to be made about what is to be tested and how extensively it is to be tested. The procedure hreaks away from the traditional "test now, decide later" schemes that have received recent criticism (e.g., Green, 1969).

A decision about mastery of one objective can be made by using the sequential probability ratio (Wald, 1947). An example of the situation is shown in Figure 5. The test length varies from pupil to pupil. A pupil is given only as many randomly-selected test items as are necessary to make a mastery or non-mastery decision with respect to a fixed mastery criterion and with prespecified Type I and Type II error rates. After each item is administered and scored, a decision is made to declare mastery, continue testing, or to declare non-mastery. With the number of items a random variable, it is possible, in this example, to make a mastery decision with as few as 6 items and a non-mastery decision with as few as 2 items. Not all mastery and non-mastery decisions are made this quickly; it depends on the response pattern of the pupil.

Figure 5 illustrates the procedure for one objective. The problem that remains is that a decision needs to be made about every objective. Since the objectives are organized into a prerequisite sequence, the sequence itself can be used in the decision-making process. This results in the compound *branching rule* shown in Table 1 for determining the next objective to be tested. The "next objective to be tested" depends on whether the student is declared a master or a non-master *and* on his response pattern that led to this decision. This is illustrated by the arrows sketched on Figure 6.

Table 1. Branching Rules for Computer-Assisted Placement Testing

Decision for 1 Skill	Pupil's Response Data (p)	Branching Rules (Next Skill to be Tested)
Mastery ($p \geq .85$)	HIGH ($p \leq .93$)	Branch **up** to **highest** untested skill.
	LOW ($.85 \leq p \leq .93$)	Branch **up** to skill **midway** between this skill and highest untested skill.
Non-Mastery ($p \leq .60$)	HIGH ($.43 \leq p \leq .60$)	Branch **down** to skill **midway** between this skill and lowest untested skill.
	LOW ($p \leq .43$)	Branch **down** to **lowest** untested skill.

Testing begins at an objective in the middle of the hierarchy and continues until the branching rule cannot be satisfied. At that point, the objective tested is the proper location of the student in the hierarchy.

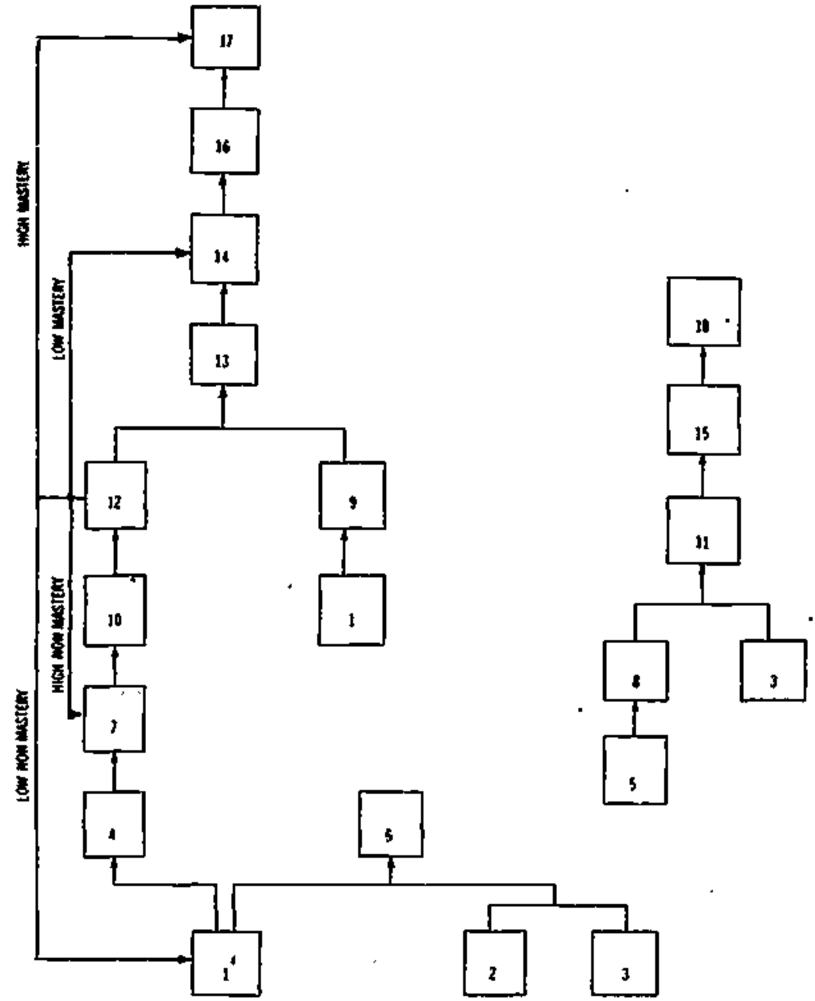


Figure 6. An Example of the Application of the Branching Rules of Table 1 to the IPI Mathematics Unit in One Instance

(Note: Only one of the "arrows" would be followed to locate the next objective to be tested. The branching rules would be reapplied after testing the next objective.)

Untested skills can be assumed mastered or unmastered according to their position in the hierarchy and the student's response data.

An individual's testing session results in a profile similar to the one shown earlier in Figure 3. The student would begin his instruction in this unit on the next sequential objective that was unmastered.

Elaborations on how items are selected and generated from item forms by the computer are given elsewhere (Ferguson & Hsu, 1971; Hsu &

Carlson, 1972). Figure 7 is a flow chart that illustrates the item selection, administration, scoring, and decision-making procedures in the testing situation. It should be noted that this type of criterion-referenced branched testing is still in the developmental stage and that evidence concerning its appropriateness needs to be provided before it can be strongly recommended.

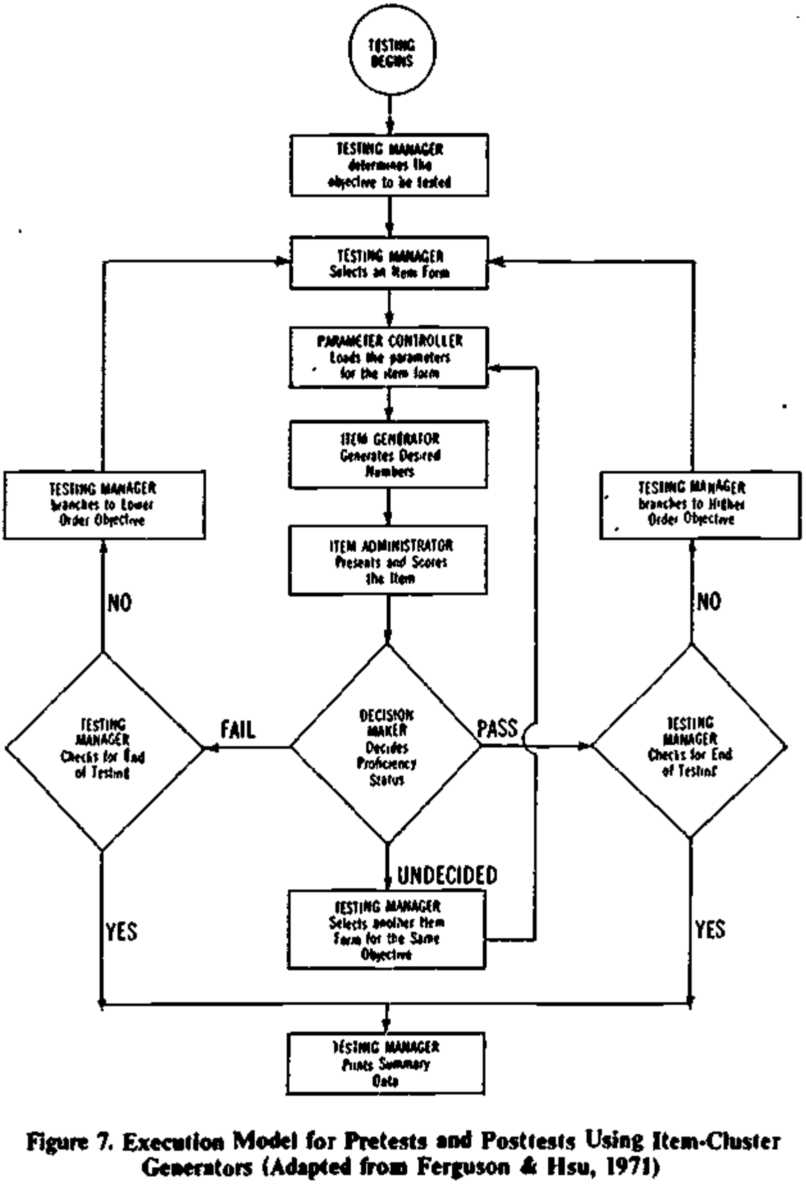

TESTING BEGINS
TESTING MANAGER determines the objective to be tested
TESTING MANAGER Selects an Item Form
PARAMETER CONTROLLER Loads the parameters for the item form
ITEM GENERATOR Generates Desired Numbers
ITEM ADMINISTRATOR Presents and Scores the Item
TESTING MANAGER branches to Lower Order Objective
TESING MANAGER branches to Higher Order Objective
NO
NO
TESTING MANAGER Checks for End of Testing
FAIL
DECISION MAKER Decides Proficiency Status
PASS
TESTING MANAGER Checks for End of Testing
UNDECIDED
TESTING MANAGER Selects another Item Form for the Same Objective
YES
YES
TESTING MANAGER Prints Summary Data

Figure 7. Execution Model for Pretests and Posttests Using Item-Cluster Generators (Adapted from Ferguson & Hsu, 1971)

CRITERION-REFERENCED TEST SCORES

Criterion-referenced test scores lead to an inference about the performance characteristics of the examinee. Such scores indicate the behaviors the examinee can exhibit with respect to a defined domain of behaviors. These scores are derived scores in the sense that their interpretation is based on the psychological structure underlying the behavior domain.

In the examples illustrated in figures 2 and 3, the unit of instruction and the node in the hierarchy are defined by classes of behaviors. A particular score on the geometry subtest, for example, might mean that the examinee can perform all lower-level behaviors up to and including: identifying pictures of open continuous curves, lines, line segments, and rays; stating how these are related to each other; writing symbolic names for specific illustrations of them; identifying pictures of intersecting and non-intersecting lines; and naming points of intersection. The score would also mean that the examinee could *not* demonstrate higher-level behaviors.

Scores may also be related to expectancy tables, thus indicating the probabilities associated with various score-behavior class performance combinations (Cronbach, 1970). This would combine norm-group data with performance data and aid in the overall interpretation of performance not tested. For example, relating acquired levels of performance to chances of being successful in new instructional situations broadens the interpretation of criterion-referenced scores. Obviously, normed-referenced scores such as percentile ranks, standard scores, grade equivalents, and so on can be obtained from criterion-referenced tests as well.

An issue often closely associated with criterion-referenced testing is that of mastery learning and mastery testing. A full discussion of mastery testing is beyond the scope of this paper. The reader is referred to papers by Bloom, Hastings, and Madaus (1971), Block (1972), Bormuth (1971), Ebel (1970), and Glaser and Nitko (1971), for some discussion of this problem as it relates to testing. It is noted here that a criterion-referenced test does not necessarily imply flawless performance nor that any examinee necessarily meet a given standard of competence. What is implied, however, is the notion that such levels of competency be defined in terms of performance (Nitko, 1970).

INSTRUCTIONAL SYSTEMS AND TESTING

It is important to point out that the kinds of tests that are developed and used will depend on the decision framework within which the test-provided information is employed (e.g., Cronbach & Glaser, 1965). It has been indicated that criterion-referenced tests will probably find their greatest use in instructional situations. Since there are a variety of ways in which instructional systems can be designed and operated to adapt to individual differences (Cronbach, 1967), the design of testing programs needs to take the instructional system into account. This means that various mixtures of criterion-referenced and norm-referenced test varieties will be needed depending on the particular instructional system. Thus, in the overall planning and designing of a testing program, decisions about when (and whether) criterion-referenced tests are to be used need to be made.

One example of how criterion-referenced and other types of test information can be designed into a particular kind of individualized instructional system has been given by Glaser and Nitko (1971). The discussion there indicates bow the various kinds of instructional decisions that need to be made are determined as well as the kinds of tests that need to be developed to provide this kind of information. Similar analyses of other types of instructional systems need to be made and testing programs need to be developed in the context of these analyses.

SUMMARY

This paper has reviewed the requirements for the construction of criterion-referenced tests that would be used in instructional situations. It has tried to indicate the problems faced in the practical construction of such tests and some of the techniques that have been found to be of some value in solving these problems. Adequate solutions do not exist for all of the problems raised. In particular, procedures are needed for the solution of the following problems:

1. Defining the behaviors to be taught and tested for in the instructional situation.
2. Task analysis as it relates to school-like behaviors.
3. Relationship between what is tested and the ultimate objectives of the individual and society.
4. The relationship between the behavioral domain and the domain of tasks serving as the potential item domain.
5. Specification of the domain of tasks in terms of their stimulus and response characteristics.
6. The ordering of the domain of behaviors in terms of their psychological structure.
7. Data related to the generalizability of samples of behavior to the behavioral domain.
8. Construct validation of proposed orderings of the behavioral domain.
9. The development of an item-writing theory and an item-response theory.
10. Development of procedures for determining mastery of identified behavior.

While solutions to the above problems would lead to improved criterion-referenced test construction practices, it should not be assumed that criterion-referenced information is all that is needed to make instructional decisions. Without an analysis of the kinds of instructional decisions that need to be made in a given instructional situation, discussions about tests, testing procedures, and test development tend to be fruitless.

REFERENCES

Baker, E.L. The effects of manipulated item writing constraints on the homogeneity of test items. *Journal of Educational Measurement*, 1971, 8, 305–309.

Block, J.H. Student evaluation: Toward the setting of rational, criterion-referenced performance standards. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Bloom, B.S., et al. (Eds.) *Taxonomy of educational objectives, handbook I: Cognitive domain*. New York: David McKay, 1956.

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

Bormuth, J.R. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1970.

Bormuth, J.R. Development of standards of readability: Toward a rational criterion of passage performance. Final Report, USDHEW, Project No. 9-0237. Chicago: The University of Chicago, 1971.

Bruner, J.S. *Toward a theory of instruction*. Cambridge, Mass.: The Belknap Press of Harvard University Press, 1966.

Cox, R.C., & Boston, M.E. Diagnosis of pupil achievement in the Individually Prescribed Instruction Project. Working Paper 15. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, 1967.

Cronbach, L.J. How can instruction be adapted to individual differences? In R.M. Gagné (Ed.), *Learning and individual differences*. Columbus, Ohio: Charles E. Merrill, 1967.

Cronbach, L.J. *Essentials of psychological testing*. (3rd ed.) New York: Harper and Row, 1970.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington: American Council on Education, 1971.

Cronbach, L.J., & Glaser, G.C. *Psychological tests and personnel decisions*. Urbana: University of Illinois Press, 1967.

Davis, F.B. Criterion-referenced tests. In *Testing in turmoil: A conference on problems and issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970.

Ebel, R.L. *Content-standard test scores. Educational and Psychological Measurement*, 1962, 22, 15-25.

Ebel, R.L. Some limitations of criterion-referenced measurement. In *Testing in turmoil: A conference on problems and issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970.

Ferguson, R.L. A model for computer-assisted criterion-referenced measurement. *Education*, 1970, *81*, 25-31.

Ferguson, R., & Hsu, T.C. The application of item generators for individualizing mathematics testing and instruction. Publication 1971/14. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, 1971.

Flanagan, J.C. Units, scores, and norms. In E.F. Lindquist (Ed.), *Educational measurement*. Washington: American Council on Education, 1951.

Gagné, R.M., & Paradise, N.E. Abilities and learning sets in knowledge acquisition. *Psychological Monographs*, 1961, 75.

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

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington: American Council on Education, 1971, 625-670.

Glaser, R., & Resnick, L.B. Instructional psychology. *Annual Review of Psychology*, 1972, 23, 207-276.

Green, B.F. Comments on tailored testing. In W. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. New York: Harper and Row, 1969.

Harris, M.L., & Stewart, D.M. Application of classical strategies to criterion-referenced test construction: An example. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hively, W. Preparation of a programmed course in algebra for secondary school teachers: A report to the National Science Foundation. Minnesota State Department of Education, Minnesota National Laboratory, 1966.

Hively, W. Domain-referenced achievement testing. A paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project*. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hively, W., Patterson, H.L., & Page, S. A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1963, *5*, 275-290.

Hsu, T.C., & Carlson, M. Oakleaf school project: Computer-assisted achievement testing. Technical Report. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, February, 1972.

Jackson, R. Developing criterion-referenced tests. ERIC/TM Report 1. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1971.

Lindquist, E.F. (Ed.) *Educational measurement*. Washington: American Council on Education, 1951.

Lindvall, C.M., & Bolvin, J.O. Programmed instruction in the schools: An application of programming principles in "Individually Prescribed Instruction." In P. Lange (Ed.), *Programmed Instruction*. 66th Yearbook, *Part II*. Chicago: National Society for the Study of Education, 1967, 217-254.

Livingston, S.A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26.

Lord, F.M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, 1958, *23*, 291-296.

Miller, R.B. Task description and analysis. In R.M. Gagné (Ed.), *Psychological principles in system development*. New York: Holt, Rinehart, and Winston, 1962.

Nitko, A.J. Criterion-referenced testing in the context of instruction. In *Testing in turmoil: A conference on problems and issues in educational measurement*. Greenwich, Conn.: Educational Records Bureau, 1970.

Osburn, H.G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, *28*, 95-104.

Resnick, L.C., Wang, M.C., & Kaplan, J. Behavioral analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. Monograph 2. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, December, 1970.

Simon, G.B. Comments on "Implications of criterion-referenced measurement." *Journal of Educational Measurement*, 1969, *6*, 259-260.

Stanley, J.C., & Hopkins, K.D. *Educational and psychological measurement and evaluation*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Stenner, A.J., & Webster, W.J. Educational program audit handbook. Arlington, Va.: The Institute for the Development of Educational Auditing, 1971.

Wald, A. *Sequential analysis*. New York: Wiley, 1947.

Wang, M.C., Resnick, L.B. & Boozer, R.F. The sequence of development of some early mathematics behaviors. Publication 1971/6. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center, 1971.