DOCUMENT RESUME

ED 137 341 TM 006 142

AUTHOR Denham, Carolyn H.
TITLE Score Reporting and Item Selection in Selected Criterion Referenced and Domain Referenced Tests.
PUB DATE [Apr 77]
NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, New York, April 5-7, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Classification; *Criterion Referenced Tests; Educational Objectives; *Elementary Education; *Item Analysis; Mathematics; Models; Norm Referenced Tests; Reading Tests; Scores; Test Construction; *Test Interpretation
IDENTIFIERS *Domain Referenced Tests

ABSTRACT

Twelve tests of reading and math at the elementary level are classified according to a model which makes a distinction between criterion and domain tests. Score reporting and item analysis techniques are discussed. It is argued that most objectives-referenced tests do not specify their domains sufficiently to make interpretations more general than the test items themselves. (Author)

ABSTRACT

Score Reporting and Item Selection in Selected Criterion Referenced and Domain Referenced Tests

Carolyn H. Denham

California State University, Long Beach

Twelve tests of reading and math at the elementary level are classified according to a model which makes a distinction between criterion and domain tests. Score reporting and item analysis techniques are discussed. It is argued that most objectives-referenced tests do not specify their domains sufficiently to make interpretations more general than the test items themselves.

Paper presented at the annual conference of the National Council on Measurement in Education, New York City, 1977.

Score Reporting and Item Selection in Selected Criterion Referenced and Domain Referenced Tests

Carolyn H. Denham

California State University, Long Beach

When we were busy creating the distinction between norm referenced and criterion referenced tests, we could overlook the difficulties with our definitions of criterion and domain referenced tests. Now it is time to make a distinction between criterion referenced tests and domain referenced tests. The present situation is too confusing. The following definition is an example (Glaser, 1971, p. 41):

> A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are referenced directly to this domain for each individual. Measurements are taken on representative samples of tasks drawn from this domain and such measurements are referenced directly to this domain for each individual measured.

The definition mixes two types of test interpretation. The first is an evaluative interpretation in which a score is evaluated in terms of performance standards or criteria. The second is a descriptive interpretation in which a score is evaluated in terms of the domain of tasks represented by the items on the test.

In this paper only those tests which compare the raw scores to performance standards will be called criterion referenced tests (or, more simply, criterion tests). Those tests in which scores on a representative sample of a clearly defined domain of items are used to estimate scores on the entire domain will be called domain referenced tests (or domain tests).

Of course, there may be tests which combine aspects of domain and criterion referencing. In an earlier paper Denham (1975) developed a model for test classification. The model provides for

seven categories of tests: criterion, domain, norm, and the four possible combinations of the three primary categories. (See Figure 1.) Examples of test score interpretations for each of the seven test categories are given in Table 1.

The present paper classifies selected criterion and domain tests according to the seven categories of the model for the dual purpose of testing the adequacy of the model and describing the current state of test development. If the tests fail to fit into the categories of the model, an attempt will be made to determine whether the model or the tests are in need of improvement.

Selected for review are tests or testing systems in math and reading for grades K-6 which are labeled criterion referenced, domain referenced, instruction referenced, or objectives referenced. Some are simply tests; others are testing systems from which the particular tests or items are selected for a given administration; for convenience the word test will include both tests and testing systems. Those tests which report group scores rather than individual scores are omitted from the discussion. Thus the exemplary domain testing in the MINNEMAST project (Hively et al., 1973) and other tests employing matrix sampling procedures are not discussed here.

A list by Kosecoff and Fink (1976) was the source of many of the test titles. Others were located through minor detective work. The list of tests is not intended to be exhaustive although the author has attempted to review most of the widely available tests or testing systems. Comments on the tests are not intended to serve as critiques of the individual tests since only certain aspects of the tests are discussed. Readers should look elsewhere for comprehensive critiques of each of the tests.

Test Classification and Score Reporting

In Tables 2, 3, and 4 are the classification and score reporting systems for those tests which fit into the seven categories of the model. Figure 2 illustrates the fact that six tests were classified as criterion, three as domain, two as norm + criterion, and one as domain + criterion. A test was classified criterion referenced if a criterion level were set and scores were reported as above or below a criterion. A test was considered domain referenced if its domain specifications were reasonably precise and/or its items were considered samples of a domain. A test was classified norm referenced if the test provided transformed scores such as percentiles or standard scores. Simply providing data on the performance of groups (such as that provided by the IOX Objective-Based Tests and by the EDITS Tests of Achievement on Basic Skills) was not considered sufficient to label a test norm referenced. Indeed, most criterion tests provide some kind of group data in the form of classroom, school, or district performance.

The items in a domain test may be written from item forms, defined by Hively, Patterson, and Page (1968) as rules for generating sets of test items. Hsu (1972) reports the use of item forms for some of the math tests in Individually Prescribed Instruction. Popham (1975) reports a simpler method of defining domains using amplified objectives as they are used in the Objectives-Based Tests of the Instructional Objectives Exchange (IOX). Even less structured is the system used by CAM, which was classified as domain primarily because longitudinal data are obtained through repeated samples of the items within each objective.

Not all of the tests examined fit into one of the seven categories of the model. The Diagnostic Math Inventory (CTB/McGraw-Hill, 1975) has only one item per objective. Hively (1974, p. 140) discusses the situation of a test with only one item per objective:

> ...the inference from the item score to the domain score is primitive: it only tells you about the probability that the students will respond correctly to the same item if you present it again.
>
> If you want stronger inference, you can construct more items for each objective, and then you can sample some of them and estimate the probability that the individual or group will respond correctly to the others. ... That is the only difference between a domain-referenced test and an objective-referenced test. The strength of the inference depends on the representativeness of the set of items associated with each objective.

Thus the Diagnostic Math Inventory does not fit the present definition of domain testing.

Nor was the Diagnostic Math Inventory classified as a criterion test. One could argue that there is an implied criterion of a correct response to the single item representing each objective, but such a criterion adds little to the test interpretation that could be achieved by simply examining the test itself. Indeed, the most basic interpretation of a score is simply to examine the test items. All of the categories in the model, however, are intended to refer to ways in which a raw score can be given meaning by referencing it to something outside the test: a norm group, a criterion level, or a domain description.

The Key Math Diagnostic Arithmetic Test (American Guidance Service, 1976) also depends on scores on single items for its "criterion-referenced" interpretations. Thus, like the Diagnostic Math Inventory, it was classified as neither criterion nor domain.

If norm referenced tests had been reviewed in this paper, the Key Math could have been classified as a norm test. Its method for producing the norm referenced scores, using Rasch-Wright procedures, is most sophisticated.

Another type of test which does not fit the model is the objectives referenced test in which items are keyed to objectives but the objectives are not adequately precise to serve as domain definitions. In most of these tests, even if they have more than one item per objective, the best way to interpret the scores is to examine the test items; examining the objectives may be misleading because their lack of specificity makes it appear that the test is more comprehensive than examination of the items reveals. Since the objectives in the Individual Pupil Monitoring System for Reading and Math (Houghton-Mifflin, 1973) were evidently not intended to serve as domain statements and no criterion level was set, the test was not placed into any of the categories.

Becoming popular are tailor-made tests in which the user selects objectives from a list and test items are compiled to meet the user specifications. Examples are the ORBIT system by CTB/McGraw Hill and many of the computer test banks discussed by Lippey (n.d.). Since these typically have neither domain statements nor criterion levels, they will generally not fit into the categories of the model. Also the reader should be aware that the items in such banks, particularly those not produced by test publishers, rarely undergo the same scrutiny as the unitary tests produced by test publishers.

Evaluation of the Model

It is evident that many tests do not fit the author's definitions of domain and criterion testing. Does this mean the model is inadequate? No, in the author's opinion, it reflects the state of the art in test development. The model lists ways of interpreting a score by referencing the score to something outside the test: a norm group, a criterion level, or a domain description. Many current tests are most appropriately interpreted by simple examination of their items; even if there is a list of objectives, interpreting the score in terms of the objectives may be making unwarranted generalizations since many objectives are not specific enough to describe adequately the items.

Interpretation of a test score by examining the items is a very useful procedure. However, there are two difficulties. The first problem, often exaggerated, is the need to keep the items secret. Fortunately there are many instances in which the students, teachers, administrators, or parents may examine the items after a test administration. In other instances, such as with a large bank of items, the items may be examined before the test administration.

The second problem is more serious. It is the fact that test developers and users want to make statements at a higher level of generality than the test itself. This is what makes the theory of educational and psychological measurement more complex than that of physical measurement. It is for the task of making generalizations from specific items that the procedure of domain testing is most promising.

Item Analysis

The model suggests a need for different item analysis procedures for each of the test categories. The first steps in an item analysis procedure can be similar for all types of tests. Whether a test is norm criterion, or domain, the items must be free of faults such as those listed in tests and measurements books. Computing difficulty and discrimination indexes and discussing the items with students are among the methods which can detect faulty items. A second measure of an item is its content validity. A test developer may consult experts for judgments about the appropriatness of each item, or the developer may use empirical techniques such as examining intercorrelations among items measuring the same objective.

Finally, the developer must select among those items which are well-written and appropriate in their content; such selection is usually necessary since there are practical limitations on the length of the test. The most efficient way of testing is to select items which contribute the most to the type of score to be reported. It is for efficiency that medium difficulty items are selected for norm tests. Survey research techniques may improve efficiency when sampling from a domain; for example, content areas in which the measurements may be less reliable can be oversampled and those areas in which correlations among items are higher can be undersampled. Efficiency of criterion tests might be improved by concentrating on items near the difficulty level of the criterion, particularly in those tests which can be scaled according to difficulty level. Efficiency on any test could be increased if those items which are

most cost-effective in terms of time are selected; this would take into consideration the fact that some types of test items take more of the subject's time than do short items such as true/false items. Other suggestions for item analysis for criterion and domain tests can be found in Denham (1975).

Item analysis procedures for each of the twelve tests were obtained through study of published manuals and through correspondence and conversations with the test publishers or developers. Item analysis information was available for all of the tests except one. The following are the findings:

1) Tests falling into different categories of the model did not have distinguishable patterns of item analysis procedures.

2) Most of the tests developers arranged for the items to be reviewed by experts in addition to empirical procedures, if any.

3) Item analysis techniques which might reveal faults in item-writing were rarely used. This was to be expected since many of the writings on criterion and domain testing disparage item discrimination and difficulty indexes although they can be quite useful in detecting item faults. In the manuals of two of the tests, however, methods for detecting poorly written items were discussed separately from other item analysis steps.

4) Some tests experimented with item analysis techniques not usually employed with norm tests. Among these techniques were sensitivity to instruction, discrimination among mastery and nonmastery groups, and a variety of procedures for evaluating the difficulty level of the items. In one test for grades 4-6 only those items which were easy for the sixth graders were chosen. In three tests, items with similar levels of difficulty were chosen to represent an objective. In two tests, items with varying levels of difficulty were chosen to represent an objective. In another test, items were chosen such that they were neither "too hard nor too easy" for the tryout group.

5) The item analysis procedure for one test consisted of administering the test to a few students and discussing the items with them.

6) The item analysis procedure of another test consisted of administering the test in one classroom to determine if all items measuring the same objective produced similar results.

7) In one test, items which fit the Rasch-Wright model were selected.

8) In one test, item forms, rather than the items, were subjected to analysis.

In summary, an impressive variety of techniques was employed. However, there was scant attention paid in most of the reports of item analyses to the detection of possible item faults. Additionally, the methods of item selection used by some of the tests were almost the opposite of those used by other tests in the same category. For example, some tests sought uniform difficulty levels; others sought variety in difficulty levels. It is almost impossible to evaluate these selection methods without a systemmatic method of determining the purposes for which the item analysis was used.

To help clear up the confusion, it is proposed that all criterion and domain tests perform each of three kinds of item analysis procedures:

1) An examination of the accuracy of the items, the extent to which the items are free of items writing faults such as those listed in tests and measurements textbooks.

2) An examination of the content of the items, the extent to which the items are representative of the objectives or the domain.

3) An examination of the efficiency of the items, the extent to which the items contribute information to the criterion or domain decision.

Attention to each of these three types of item analysis data would mean that criterion and domain test developers would no longer neglect examination of item accuracy. It should also help test developers think more clearly about the purposes of their item analysis procedures and to invent new methods of item analysis, whether empirical or judgmental.

Other Technical Considerations

Perfecting item analysis techniques for domain and criterion tests is only one of the many tasks remaining for researchers. The issues of reliability, validity, estimation of domain scores, and estimation of mastery states are among the current problems.

Livingston (1972), Huynh (1976), and Swaminathan (1974) report methods of computing reliability applicable to criterion testing. We are in need of research on methods of computing reliability of domain estimates.

Maskauskas (1976) discussed at length the problem of cut-off scores for criterion tests. We are in need of research on the estimation of domain scores from test items. Although the problem may seem a simple one, it is actually quite complex. Two of the models which have been proposed for estimation of domain scores are the binomial model (Millman, 1974), in which the percentage of items answered correctly on the test is taken as a point estimate of the domain score and group data is not considered, and the classical testing model, in which the estimated domain score is a regressed score utilizing data on group performance. These two models were criticized by Haladyna (1975). Another procedure for estimating domain scores is the Bayesian approach in which group data or other data may be used as prior information (Lewis et al., 1973 & Novick et al., 1973). The Rasch-Wright model and Cronbach's theory of generalizability are two other models which could provide domain score estimates.

Summary

The paper describes twelve criterion and domain tests in terms of a model which makes distinctions between criterion and domain tests. Score reporting and item analysis procedures are discussed. The author found few tests which could legitimately be called domain and many objectives referenced tests which did not fit the model. This reveals a problem in current test development; although many tests are interpreted in terms of performance of objectives, these objectives are often too loosely written to serve as descriptions of the actual items or, on the other hand, the items are too few and too homogeneous to represent the more broadly stated objectives. In place of objectives referenced tests the author advocates domain referenced tests in which the domains are clearly specified and an attempt is made to choose items which are representative of the domains.

Of course, if item forms as complex as those written by Hively et al. (1973) must be used, many test developers might simply refuse to try. Specification of and sampling from domains is a matter of degree. The present author recommends more careful attention to specification and sampling so that one may interpret the test at a higher level of generality than the test itself. However, it is hoped that test developers do not make the task so complex that they lose themselves in their domains. The English essayist Charles Lamb (1823) must have been referring to such a situation in his account of one man: "He was lord of his library, and seldom cared for looking out beyond his domains."

REFERENCES

Denham, C.H. Criterion-referenced, domain-referenced and norm-referenced measurement: a parallax view. Educational Technology, 1975, 15, 9-13.

Glaser, R. A criterion-referenced test. In Popham, W.J. (Ed.), Criterion-referenced measurement; an introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Gorth, W.P., O'Reilly, R.P., & Pinsky, P.D. Comprehensive Achievement Monitoring: A criterion-referenced evaluation system. Englewood Cliffs, N.J.: Educational Technology Publications, 1975.

Haladyna, T. An analysis of two procedures for decisionmaking when using domain-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., April 1975.

Hively, W. (Ed.) Domain-referenced testing. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: a technical handbook and a case study. In Alkin, M.C. (Ed.), CSE monograph series in evaluation. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Hively, W., Patterson, H.L., & Page, S.H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Hsu, T. & Carlson, M. Oakleaf School Project: Computer-assisted achievement testing. Learning Research and Development Center, University of Pittsburg, 1972.

Huynh, Huynh. On the reliability of decision in domain-referenced testing. Journal of Educational Measurement, 1976, 4, 253-563.

Kosecoff, J. & Fink, A. The feasibility of using criterion-referenced tests for large-scale evaluations. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1976.

Lamb, Charles. Poor Relations, 1823.

Lewis, C., Wang, M., & Novick, M.R. Marginal distributions for the estimated proportions in m groups. Technical Bulletin, No. 13. Iowa City: American College Testing Program, 1973.

Lippey, G. Summary of computer-assisted test construction systems. San Jose, CA: Classroom Support Systems, n.d.

Livingston, S. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 1, 13-26.

Meskauskas, J.A. Evaluation models for criterion-referenced testing: views regarding mastery and standard setting. Review of Educational Research, 1976, 46, 133-158.

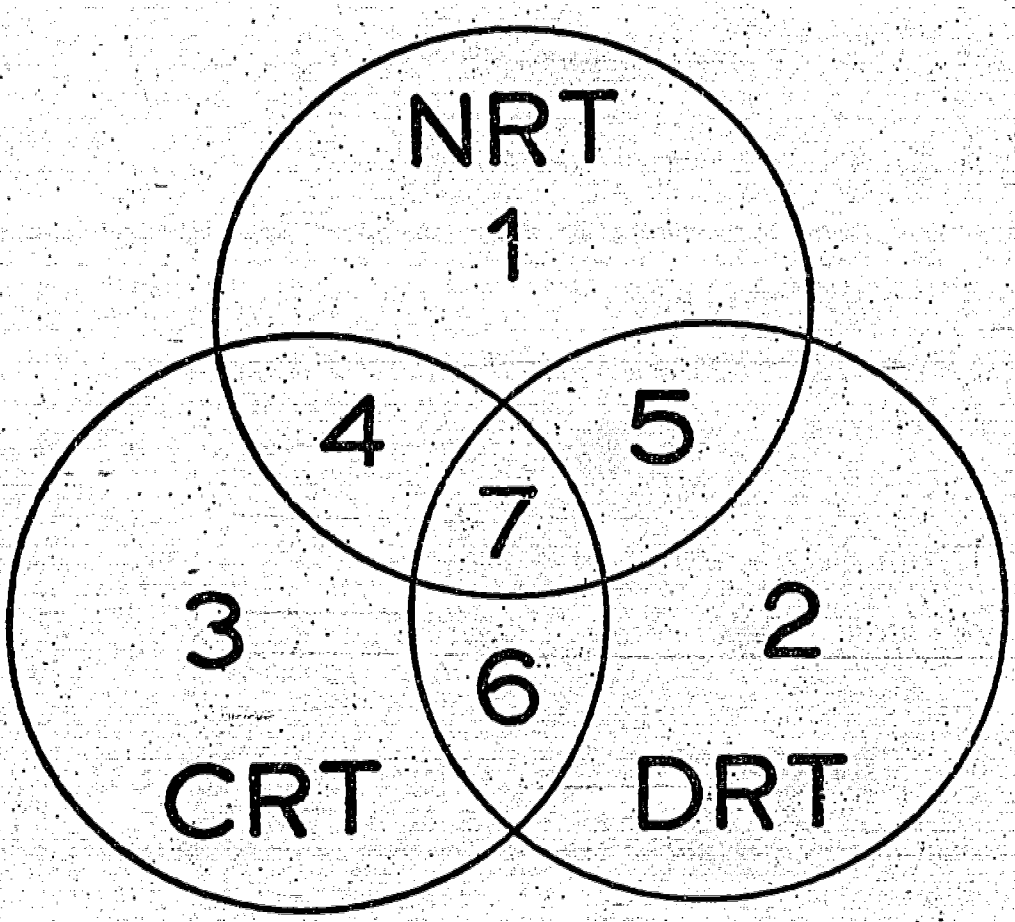
Novick, N.R., Lewis, C., & Jackson, P.H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-46.

Popham, W.J. Educational Evaluation. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1975.

Sension, D.B. & Rabehl, G.J. Test item domains and instructional accountability. In Hively, W. (Ed.), Domain-referenced testing. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Swaminathan, H. Hambleton, R.K., & Algina, J.J. Reliability of criterion-referenced tests: a decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-268.

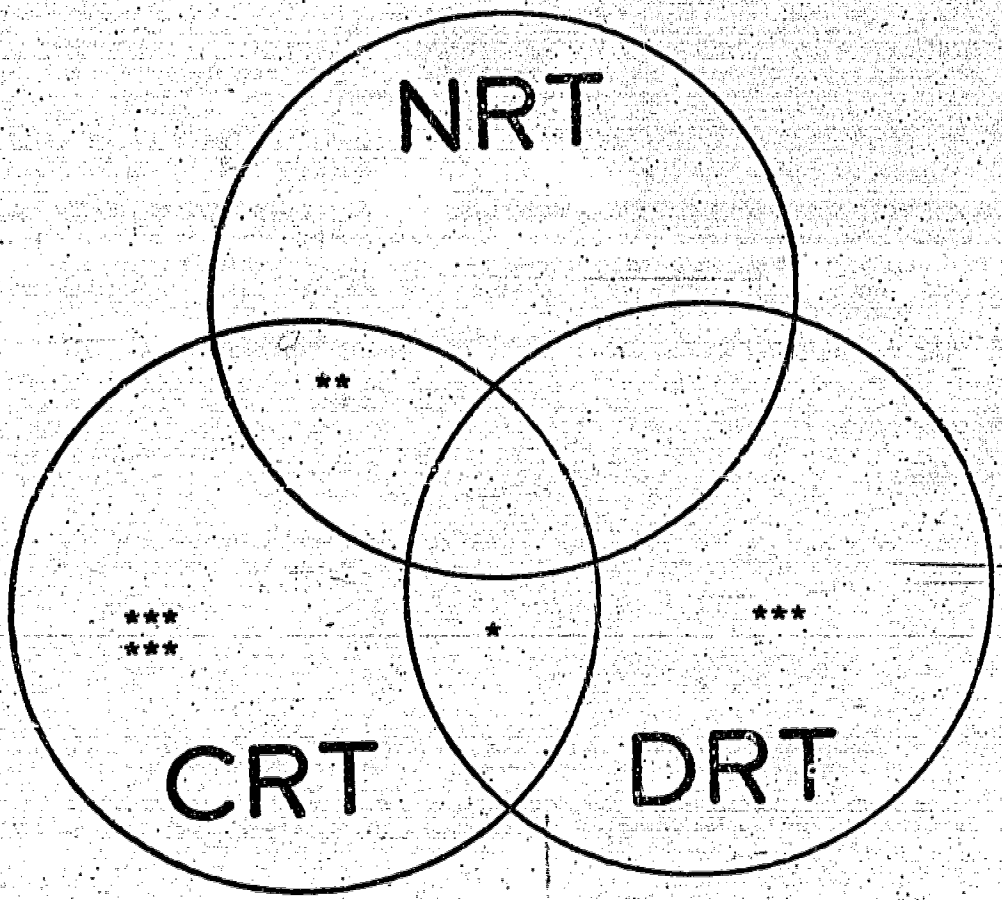
FIGURE 1



Seven categories of tests:

Norm referenced, criterion referenced, domain referenced and the 4 possible combinations.

FIGURE 2



The placement of twelve reading and math tests within the model.

Table 1

Description of Score Reporting for Each of the Seven Categories of Tests

1) NR: The student scores better than 85% of the norm group.

2) DR: The student correctly spelled 9 out of 10 words randomly chosen from the list of sixth grade spelling words. It is estimated that he can spell 90% of the words on the list.

3) CR: The student met the criterion of 80% of the words spelled correctly.

4) NR + CR: The student scored better than 85% of the norm group and met the criterion of 80% of the words spelled correctly.

5) NR + DR: It is estimated that the student would score better than 85% of the norm group on the entire list of sixth grade spelling words. It is estimated that the student can spell 90% of the words on the list.

6) CR + DR: It is estimated that the student would meet the criterion of 80% correct on the entire list of sixth grade spelling words. It is estimated that the student can spell 90% of the words on the list.

7) NR + DR + CR: It is estimated that the student would score better than 85% of the norm group and would meet the criterion of 90% correct on the entire list of sixth grade spelling words. It is estimated that the student can spell 90% of the words on the list.

Table 2 Score Reporting on Criterion Tests

Test, Publisher, Date	Test Category	Score Reporting
Skills Monitoring System - Reading The Psychological Corporation/ Harcourt Brace Jovanovich, Inc. 1974-75	CRITERION	On the Skill Locator (a survey) there are two items per objective. Both must be answered correctly for mastery. On the shorter Skill Minis (8-12 items), 80% must be answered correctly for mastery.
Mastery: An Evaluation Tool SOBAR Reading and Mathematics Science Research Associates 1975	CRITERION	The mastery tests contain three items per objective. All three must be answered correctly for mastery.
Fountain Valley Teacher Support System - - Reading and Mathematics	CRITERION	Seventy-five percent of the items must be answered correctly for proficiency on an objective. Each test measures approximately six objectives.
Prescriptive Reading Inventory CTB/McGraw Hill 1972	CRITERION	Number of items per objective varies. In the case of three items, two out of three correct indicates mastery. For four items, three out of four indicates mastery.
Tests of Achievement in Basic Skills -- Mathematics and Reading Educational and Industrial Testing Service	CRITERION	Criterion varies according to subject matter and level. In Level 2 reading, objectives with three items or less require 100% proficiency. The criterion level is 75% for objectives with four or more test items. Level B Math has only one item per objective. A correct response to the item indicates accomplishment of the objective.
Doren Diagnostic Reading Test of Word Recognition Skills American Guidance Service 1973	CRITERION	If more than six items are answered incorrectly in any skill area, remediation is indicated for the area. This is equivalent to a criterion of 70%.

Table 3

Score Reporting on Criterion Tests

Test, Publisher, Date	Test Category	Score Reporting
Objectives-Based Tests Instructional Objectives Exchange (IOX) 1974	DOMAIN	Number correct is reported for each amplified objective. There are 5 or 10 items per amplified objective. Some normative data is available but raw scores are not converted to normative scores.
Reading Block Assessment and Reading Placement Aid SWRL Educational Research and Development/Ginn & Co. 1976	DOMAIN	On each of the eight Block Assessments, the number of items correct on each of four outcomes is reported. However, the Reading Placement Aid is criterion referenced. The first page on which the pupil scores 6 or less determines the suggested block assignment.
Comprehensive Achievement Monitoring (CAM) Systems	DOMAIN	Implementation varies from system to system, but those in which items are sampled from domains may be called domain tests. Scores are reported for each objective. Typically the objective is tested at several different points in time using different samples of items to measure the objectives, producing longitudinal data. See Gorth et al. (1975) and Sension and Rabehl (1974) for more information.