DOCUMENT RESUME

ED 084 295 TM 003 309

AUTHOR	Fremer, John
TITLE	Application of Criterion-Referencing to Schools.
NOTE	18p.; Paper based on a presentation at the Annual Conference of the Educational Records Bureau (N.Y., N.Y. November 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Criterion Referenced Tests; *Data Collection; Measurement Techniques; Norm Referenced Tests; *Student Evaluation; Technical Reports; *Test Construction

ABSTRACT

The definitions of norm-reference" and "criterion-referenced" tests and the applications of them are discussed. A norm-referenced test is one that yields scores which are interpreted through the use of norms. Applications of these include survey achievement tests, college selection tests, aptitude batteries, and armed services classification tests. The word "criterion" has been used with two different meanings. Psychologists and educational researchers use the term as a specific standard against which learning performance is judged. Measurement people use the word to refer to prformance in the real world against which test performance will be validated. There are also differences in approach to the word "referencing." It is suggested that criterion-referenced tests can be differentiated from norm-referenced tests in that they do not focus on the problem of individual differences. The problem of developing criterion-referenced tests is discussed. An assessment program is an appropriate place to apply criterion-referenced concepts and techniques, as specific information on students' attainment of objectives can be collected. (CK)

Applications of Criterion-Referencing to Schools[1]

John Fremer[2]
Educational Testing Service

Interest in criterion-referenced testing is at a very high level now, yet it seems to be growing rapidly with each passing month. As part of my job at Educational Testing Service, I have attempted to keep track of activities in the area, and at one point a few months ago I thought I was really on top of what was happening. I was involved in a number of criterion-referenced testing projects at ETS, I had visited schools that had developed criterion-referenced testing programs, and I had collected and read some 200 articles and papers on the subject. In the last month, however, I learned that the ERIC Tests and Measurement Center has identified and annotated some 300 articles on the topic. When they turn out a report incorporating these annotations, it will be a valuable resource to all of us who are interested in applying criterion-referenced testing concepts to schools.

Definitions

One of the conclusions I have come to on the basis of working in the area of criterion-referencing is that there is a great deal of confusion about the meaning of the term. Before I attempt to review possible applications, therefore, I am going to consider the two terms -- "norm-referenced" and "criterion-referenced."

[1] Paper based on a presentation at the Annual Conference of the Educational Records Bureau, New York City, November, 1973.

[2] Associate Director of Elementary and Secondary School Programs at Educational Testing Service.

Norm-Referenced

I will start with "norm-referenced." A norm-referenced test is one that yields scores which are interpreted through the use of norms. Scores are typically referred to one or more tables of norms. The tables supply for any given score a percentile standing with respect to some population or reference group. Norm-referenced tests rely on comparisons of performance to bring meaning to individual scores. They usually do not provide information on the specifics of student competencies. Some examples of norm-referenced tests are the following:

1. Survey Achievement Tests -
   Iowa Tests of Basic Skills
   Stanford Achievement Tests
   Sequential Tests of Educational Progress

2. College Selection Tests
   Admissions Testing Program of the College Board -
   Scholastic Aptitude Test and the Achievement Tests
   American College Testing Program

3. Aptitude Batteries
   Differential Aptitude Tests

4. Armed Services Classification Tests

It is perhaps worth noting, though, that some of the applications of the norm-referenced tests that I used as examples involve the use of test performance standards. Specific scores are identified as being the minimal acceptable ones, so that individuals earning scores at that level or above, are said to have met the criterion and individuals earning scores below that level are considered to have failed to meet criterion. The word "criterion" is used in this context to mean "standard."

It's quite common for people to use the mean as a standard and to ask that schools bring everyone up to the mean.

Criterion-Referenced

Let's turn to the term "criterion-referenced." One of the problems with this term is that the word "criterion" has been widely used in different contexts that are somewhat independent of its use in the term criterion-referenced. The word "criterion" has been used, moreover, with two different meanings.

Psychologists and educational researchers have tended to emphasize the meaning of the word "criterion" that is employed in learning experiments. The criterion in that sense is the level of performance or "score" that is taken to mean that the subject or student has learned. The criterion is thus a specific standard against which learning performance is judged.

Measurement people, on the other hand, have tended to use the word criterion to refer to behavior or performance in the real world against which test performance will be validated. The test makers have talked about validating college selection tests, for example, as predictors of college success. The behavior "success in college" is a complex one, not one that is precisely defined or directly observable. What testmakers have tended to use as measures of the criterion behavior "success in college" is grade point average (GPA) or, less often, satisfaction scales. Somewhat the same situation holds for tests designed to predict job success or satisfaction. The criterion is some complex set of behaviors that are measured through some indirect process.

We have a situation, therefore, where the word "criterion" is used by some people to mean a standard and by others to mean some important real-life behavior.

Not only are there somewhat different points of view with respect to the word "criterion", there are also some differences in approach to the word "referencing."

Some writers and practitioners in the area of criterion-referencing have focused their attention on technical procedures for relating test behavior to classroom or other real-world behavior. (They have emphasized methods of scaling or of validating inferences from tests.) Other measurement and educational workers have devoted their energy to the careful specification of areas or domains of test coverage. Still others have concentrated on the role of individual objectives with less attention to the relationship among these objectives.

With my comments about the words "criterion" and "referencing" as background, let's look at some definitions. If you were to turn to the Thorndike-edited book <u>Educational Measurement</u>, to the chapter by Glaser & Nitko, you would find the following definition of a criterion-referenced test:

> "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards."
> (Glaser & Nitko, 1971, p. 653)

In the discussion which follows this definition in the Glaser & Nitko chapter, it is suggested that criterion-referenced tests can be differentiated from norm-referenced tests in that they do not focus on the prohlem of individual differences. Criterion-referenced tests are not designed to determine an individual's relative standing in some norm group. <u>Rather,</u>

they tell you what an individual can or cannot do. Glaser & Nitko talk about the need to construct a criterion-referenced test by defining a population of tasks, such as all possib'e pairs of two-digit numbers that might be added or a specified list of words all of which would have to be correctly spelled. These examples focus on knowledge or skills, but the same logic could be applied to non-cognitive or affective areas.

The Glaser & Nitko definition of a criterion-referenced test has received considerable attention, but other definitions are also available. Robert L. Ebel characterizes criterion-referenced measurement as follows:

> "The essential difference between norm-referenced and criterion-referenced measurements, is in the quantitative scales used to express how much the individual can do. In norm-referenced measurement the scale is usually anchored in the middle, on some average level of performance for a particular group of individuals. The units on the scale are usually a function of the distribution of performances above and below the average level. In criterion-referenced measurement the scale is usually anchored at the extremities, a score at the top of the scale indicating complete or perfect mastery of some defined abilities, one at the bottom indicating complete absence of those abilities. The scale units consist of subdivisions of these total score ranges." (Ebel, 1971, p. 282)

Notice that Ebel talks about criterion-referenced "measurement" or the entire process of which the test is one part. In this connection, Ebel calls our attention to the scale used for expressing results. Glaser and Nitko define a criterion-referenced "test" and indicate that it is the method of construction that makes possible the criterion-referenced scale that Ebel describes. Both these definitions agree that a criterion-referenced scale is one which permits direct interpretation with regard to performance standards, but the definitions differ in the requirement that the interpretation depends on the method of construction for the criterion-referenced test. Ebel's definition would be satisfied if a test was used in a validation study that determined experimentally which inferences regarding a student's non-test behavior could be legitimately made from given test scores.

Perhaps these two definitions could permit us to consider a general definition such as the following:

> A criterion-referenced test tells you something about a person without reference to the performance of any other person. In a cognitive area, it tells you what a person knows or can do. In an affective area, it indicates a person's attitudes, values, preferences, etc.
>
> A criterion-referenced test assumes that behavior can be measured in such a way that comparison with a standard is possible.

Applications

There is certainly more that could be said about definitions. Perhaps I have said enough, though, to move on now to the issue of applications.

Generally speaking, you can use criterion-referenced tests for just about any of the purposes for which you have used norm-referenced tests. In fact, the same tests can be used in either a criterion-referenced or a norm-referenced way. The crucial distinction lies in whether you rely on norms for your score interpretation or whether you develop your tests and your program so that comparisons with standards are possible.

Consider the use of tests for selection, for example. There may be situations where you would want to have something more than a global test score when you make selection decisions. You may want a test that provides indications of specific competencies. This same reasoning would hold for the use of tests for placement.

The amount of selection or placement information you can get from test results will depend greatly on the amount and type of testing that you can do, and the clarity with which you can formulate objectives and develop associated items or exercises.

Most of my own experience with criterion-referenced and objectives-referenced tests has not been with selection and placement uses but with the use of such tests for either assessment or instructional management. It is on these two areas that I will focus the bulk of my remaining remarks.

I like to approach the problem of helping people develop criterion-referenced tests for either assessment or instructional management by asking them to consider the kinds of reports on student performances that are desired. What information is needed, who will use it, what decisions will be made on the basis of the results? These are the kinds of questions that need to be answered before much progress can be made with planning.

Assessment -- Using Existing Tests

When you plan an assessment program you need always to decide whether to use existing tests or to develop your own tests. Let's start with the possibility that you have identified an existing test that is close enough a match to your existing program in a particular subject matter area for you to feel comfortable using it in a school assessment program.

Most existing achievement tests that are available to schools are survey tests that contain relatively few items per objective. This is true regardless of the label that may appear on the tests. Although these tests are designed for normative interpretations, it is possible to set criterion levels or cutoff scores for the tests on some more logical and systematic basis than that of calling the mean the "standard" that is to be achieved. One way of setting such a cutting score or criterion level is to ask teachers to evaluate the individual items in a test and to estimate what the difficulty of each item should be for a student at some competency level that is of interest. To do this job you need first to develop a comprehensive definition of the competencies you are interested in, possibly including specification of the subsequent educational experience for which the student needs to be prepared. The items in a mathematics test for sixth graders, for example, might be evaluated in terms of their estimated difficulty for those sixth grade students who are considered sufficiently well trained to be ready for an early algebra course or some other further mathematics instruction.

This type of estimating is possible and has been used by ETS as a method for setting test cut-off scores or criterion levels that will correspond to some carefully defined competency level. It requires the judges, however, to carry out an extremely difficult assignment. They have to develop a concept of a person who just meets the required competency level and then evaluate items in terms of such a hypothetical person. When carrying out such a task, the availability of data on student performance on items related to the required skills seems almost essential. Ideally, data would be obtained on students who are identified by methods independent of the test, as being above or below minimum competency. A criterion level can then be set that differentiates between these two groups.

This procedure can be applied with either a total test or with clusters of items from a test that measure the same or similar objectives. In general, the more variety there is in the coverage of the test the more global will be your information about student competencies. When you are interested in group assessment, though, rather than individual diagnosis, you can use global test information to make good inferences about specific competencies. Total scores on a survey reading test, for example, could be used to estimate what proportion of a group have mastered certain reading skills or what proportion could read a particular passage and to report or recognize the main idea. This notion of using existing survey achievement tests to make inferences about how groups stand with respect to specific competencies is developed in a report of mine entitled "Criterion-Referenced Interpretations of Survey Achievement Tests" (Fremer, 1972). As an appendix to this paper I have listed this paper and some other articles that seem to me to provide gcod ideas for people interested in criterion-referencing.

Assessment -- Developing New Tests

I have discussed the use of survey achievement tests in an assessment program to make criterion-referenced inferences because this seems to me to be a possibilitv that has not received much attention in testing journals. I believe, however, that most measurement people would recommend a locally tailored criterion-referenced assessment program when a school or school district can afford to commit the necessary resources to the job. I have identified several aspects of the development of such a program that will require attention.

First, and clearly important, is the task of specifying overall program objectives. In most instances this issue will prove to be closely related to the problem of deciding who are the audiences for the assessment program results. Within the school setting potential audiences include:

students

teachers

curriculum specialists

counselors

educational administrators

From the larger community, parents are obviously a prime target audience, but school boards and community groups of various kinds have interests that also need to be considered. If you fail to consider any of these audiences, you will surely hear complaints. In each school setting the priority to be assigned to each audience will influence the planning and development of a criterion-referenced assessment program. The long-range program goal should be to produce information of interest and value to all of these audiences. You have to start somewhere, though, so an initial decision might be made to try to identify several groups inside and outside the school setting that will be the first to receive assessment information. The program can then be expanded in subsequent years to reach other audiences.

When you have identified the groups to be served, it will be highly desirable to work with these groups both on the identification of priority educational objectives to be assessed, and to design meaningful reporting formats. It is my impression that schools have done a fairly good job of involving the various audiences for assessment information in the jobs of developing goals and objectives and of setting priorities. Not enough attention has been given, though, to the methods of reporting assessment data. Too often reporting is an afterthought. When this happens, the data obtained may well be misinterpreted, ignored, or rejected. This is an issue that has received a great deal of attention in the development of exercises and test questions for the National Assessment of Educational Progress, ε program for which ETS has developed exercises in several areas. In the course of developing exercises for the National Assessment, reportability is a major criterion at each stage of the development procesɜ, yet neither the reviewers of National Assessment reports nor the assessment staff have been completely satisfied with any of the reports ɼroduced to date.

The educational objectives for the school will become the basis for item or test question selection or development for the criterion-referenced testing program. Although there are many moaels of educational objectives available to use as starting points for local objectives development, staff and community cooperation on a local set of objectives offers considerable advantages to a school. It can promote involvement in a program and increased awareness of the extraordinary breadth of the objectives that a school is expected to attain. From my reading of school-produced goal statements, I have been impressed by the predominant importance often assigned to such non-cognitive areas as "self-concept" and "lifelong commitment to learning." These are often valued even more highly than reading, writing, and arithmetic, even though it is these latter skills areas that are the current basic elements of almost all assessment programs.

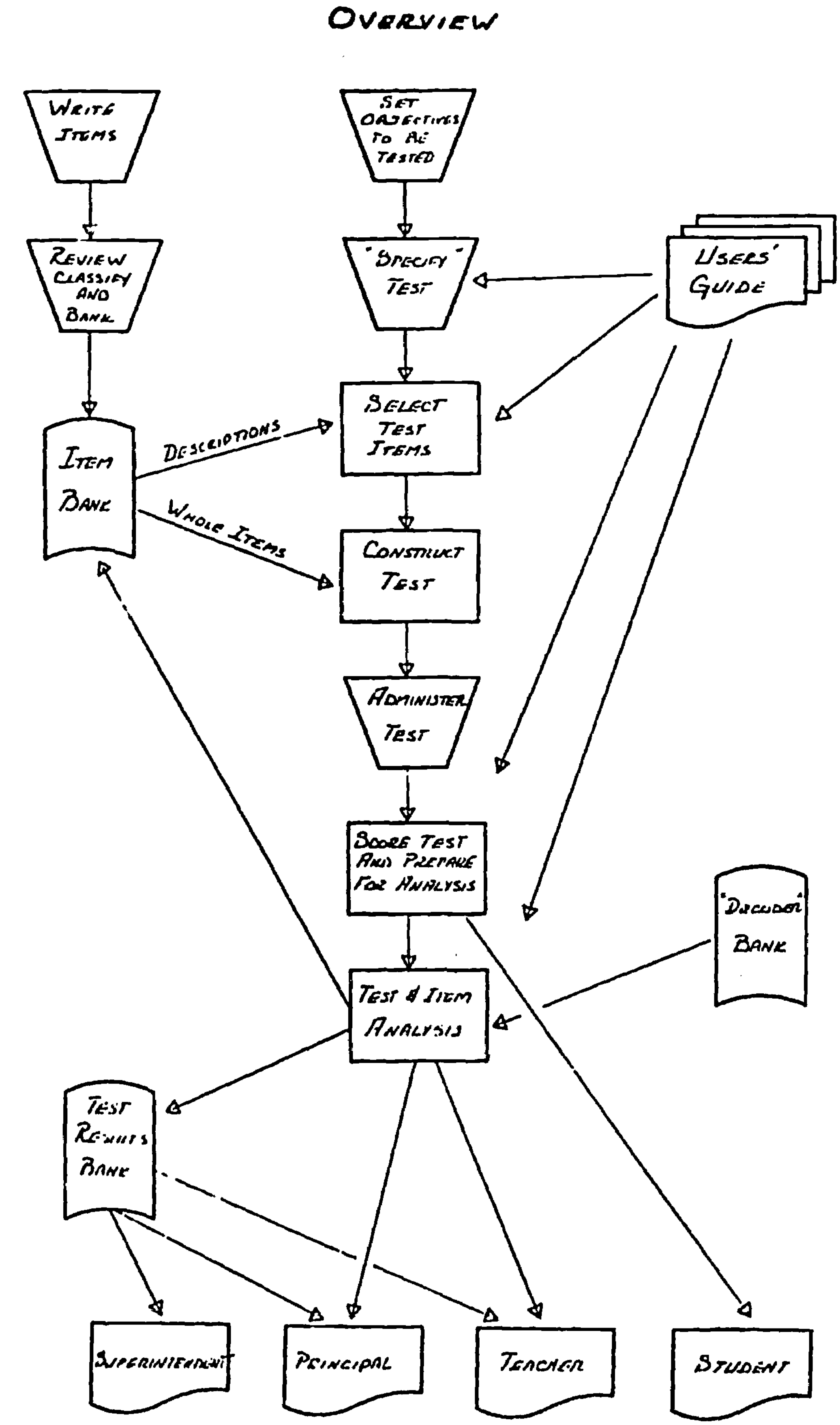
The heavy emphasis on the communications and mathematics skills areas in assessment programs probably relates closely to the fact that these areas are relatively easy ones to measure. Testing what is easiest to measure rather than what is most valued is a problem with criterion-referenced testing, just as it has been a problem with norm-referenced testing. The available survey achievement tests of the test publishers and items in the public domain such as those released by the National Assessment are heavily focused on skills, as opposed to noncognitive areas. The district or school that hopes to produce a balanced assessment program will need to face difficult test development problems in order to han.'le objectives relating to areas such as "self-concept." If the easy, skills-only route is taken at the start of an assessment program, though, it will be difficult to get up the courage to expand into the non-cognitive area at a later time.

Figure 1 is a generalized flowchart of a criterion-referenced testing system for a school or school district.

Fig. 1

Prepared by Jesse Webb, ETS, 1973

Simple Criterion-Referenced and Objectives-Referenced Testing System



This flowchart assumes that the purposes for a program have been established and that lists of educational objectives have been developed. These objectives serve as the basis for the development of test items which are then reviewed and stored in an item bank. The item bank may well be a set of file drawers in which items are stored behind index cards corresponding to objectives. Alternatively the items could be stored on looseleaf type notebook pages with each page containing the items tied to an objective. This model system assumes that teachers or administrators will make their first decision about a test in terms of the objectives that will be covered; so the first step in test development is labeled -- "Set Objectives to be Tested." Next, the system user makes decisions about the number of items to be administered and other related decisions that complete the specifications or requirements for the test that is needed.

The subsequent steps in the outline take the specifications and use them to develop tests which are administered, scored, analyzed, and reported. This outline shows only in-school audiences for reports, i.e., superintendent, principal, teachers, and student. As I indicated earlier, the number of potential audiences is much larger.

Notice that the flowchart calls attention to the need for a "Users' Guide." This guide should be developed jointly by the technical systems people involved in a criterion-referenced assessment project and by the users of the system. The guide would provide the information that the user would need to carry out each of the steps.

Notice also that the system calls for a test results bank. The value of such a bank would depend on the extent to which school people found that it was helpful to them in making decisions. It could be just a file in the corner of a school records room that no one ever bothered to look at, except to toss in another batch of test results. It could, on the other hand, be a resource for future educational program planning. Teachers and curriculum specialists might review the results for last year's classes to see which objectives were being attained and which were not. Program changes or adjustments of the age or grade level for objectives could be made.

Instructional Management

I believe that an assessment program is a very appropriate place to apply criterion-referenced concepts and techniques. Specific information on students' attainment of objectives can be collected. This information will almost certainly be valued by teachers and other school personnel if their needs were considered at every point of program planning and implementation. I also feel, however, that the most powerful application of criterion-referencing is in the area of instructional management. Some of the initial work on criterion-referenced instructional management systems dealt with the task of developing tests that would be an integral part of particular curriculum materials. Each unit in the instructional materials would be followed by a test covering the objectives of those materials. Although this continues to be an approach that some curriculum developers are following, it is not the application I plan to address. When conducting classroom lessons, teachers seem most often to combine materials from more than one packaged set of curriculum offerings and to add to this mixture their own ideas and approaches. A fixed materials and testing sequence is not likely to match what is actually going on in the classroom. What is needed is a flexible criterion-referenced testing system that is controlled by the teacher. Figure 1 can provide the general basis for such a system, but only if a continuing work team is created within a school or school system to design and carry out all the necessary steps. This group will need to take into account such constraints as the state of the school's budget and the availability of appropriately trained and motivated staff.

As in any project, the number of dollars needed will be a major limitation on what can be done. The idea of a flexible criterion-referenced instructional management system may seem completely unworkable if you assume that start-up costs have to be very substantial before there is any return. A particular school or school system may lack personnel with measurement training and may not have computer facilities. Administrators may want to avoid, therefore, the addition of new demands on limited available funds. Some positive factors can be considered, though. It would be possible to establish a prototype criterion-referenced instructional management system in a school without making the system dependent on a computer. Hand-processing procedures could be developed that could eventually be converted to computer procedures when and if the school does move to computer facilities for other aspects of school operation. A decision to computerize would only be made if the program proved to be used and valued by school staff. Initial planning of a hand-processing system would need to take into account the possibility of a future conversion to a computer system so that a transition could be accomplished without much added expense.

The problem of not having adequately trained staff to initiate and sustain a project can be addressed by calling on outside help. The use of outside experts, though, should be carried out in a way that builds the competencies of existing staff. Initial assistance might take the form of workshops on the development of objectives and on criterion-referenced testing applications. Such workshops could be part of ongoing in-service training programs. Key staff members might attend special training programs such as the Intensive Residence Course on Criterion-Referenced and Objectives-Referenced Measurement offered by the Programs of Continuing Education at Educational Testing Service. Visits to schools that have already implemented similar systems can provide models and even materials for your use.

Final Thought

One final statement that I would like to make is that "it won't be easy." I certainly believe that. I also believe, though, that the labor is justified. In their efforts to deal with students as individuals and to help them learn, teachers are continually making decisions about instruction. These decisions are made on the basis of a great deal of information, much more information than could ever be obtained from tests. Yet tests could do part of the job, certainly more than they are now doing. The application of criterion-referenced approaches can help provide information that teachers, administrators, and students can use to facilitate learning.

APPENDIX: Selected Papers on Criterion-Referencing[1]

Ebel, Robert L. Criterion-referenced measurement: limitations. *School Review*, 1971, *79*, pp. 282-288.

Fremer, John *Criterion-Referenced Interpretations of Survey Achievement Tests*. TDM-72-1, Princeton, New Jersey: Educational Testing Service, 1972.

Fremer, John Developing a criterion-referenced assessment program. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, February 1973.

Fremer, John Services in the area of criterion-referenced and objectives-referenced measurement: what, why, and where next. Paper adapted from a presentation at the Michigan School Testing Conference, Ann Arbor, Michigan, March 1973.

Glaser, Robert & Nitko, Anthony J. Measurement in learning and instruction. In Robert L. Thorndike (Ed.), *Educational Measurement*, Washington, D. C.: American Council on Education, 1971. pp. 625-670.

Hambleton, Ronald K. & Novick, Melvin R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, pp. 159-170.

Hawes, Gene R. Criterion-referenced testing: no more losers, no more norms, no more parents raising storms. *Nation's Schools*, February 1973, Volume 91, No. 2.

Hsu, Tse-Chi & M. Elizabeth Boston Criterion-referenced measurement: *An Annotated Bibliography*. University of Pittsburgh: Learning Research and Development Center, 1972.

Jackson, Rex *Developing Criterion-Referenced Tests*. TM Report No. 1, Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurements, and Evaluation, 1972.

Klein, Stephen P. and Kosecoff, Jacqueline *Issues and Procedures in the Development of Criterion-Referenced Tests*. TM Report No. 26, Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1973.

Millman, Jason Criterion-referenced measurement: an alternative. *Reading Teacher*, 1972, *26*, pp. 278-281.

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

[1] This list includes references cited in this paper.