DOCUMENT RESUME

ED 135 840 95 TM 006 067

AUTHOR Kosecoff, Jacqueline; And Others
TITLE A System for Describing and Evaluating Criterion-Referenced Tests.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ERIC-TM-57
PUB DATE Dec 76
CONTRACT 400-75-0015
NOTE 15p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Evaluation; *Evaluation Criteria; *Evaluation Methods; *Rating Scales
IDENTIFIERS Criterion Referenced Test Description Evaluation

ABSTRACT

There are, at present, a number of tests that are labeled criterion referenced. These tests vary considerably in format, design, analysis, and function. In order to provide an efficient and objective procedure for describing, assessing, and comparing these measures, the Criterion Referenced Test Description and Evaluation (CRTDE) rating system was developed. This system incorporates general concern for the overall characteristics and usability of a CRT and a specific concern for the technical excellence with which the CRT was developed and analyzed. The CRTDE rating form is divided into eight parts. The first three pertain to overall CRT characteristics and usability: (1) marketing and packaging, (2) examinee appropriateness, and (3) administrative usability. The second five pertain to CRT technical excellence: (4) function and purpose, (5) objectives development, (6) item development, (7) methods of score interpretation, and (8) analysis and validation. In addition to an explanation of the rating system, this paper includes detailed instructions so that it can be used in a standardized and accurate way by school personnel, test-selection committees, researchers, and professional evaluators. (Author/RC)

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 57 DECEMBER 1976

A SYSTEM FOR DESCRIBING AND EVALUATING CRITERION-REFERENCED TESTS*

Jacqueline Kosecoff, Arlene Fink, & Stephen P. Klein

ABSTRACT

There are, at present, a number of tests that are labeled *criterion-referenced*. These tests vary considerably in format, design, analysis, and function. In order to provide an efficient and objective procedure for describing, assessing, and comparing these measures, the Criterion-Referenced Test Description and Evaluation (CRTDE) rating system was developed. This system incorporates a general concern for the overall characteristics and usability of a CRT and a specific concern for the technical excellence with which the CRT was developed and analyzed.

The CRTDE rating form is divided into eight parts that reflect these concerns:

Overall CRT Characteristics and Usability
1. Marketing and Packaging
2. Examinee Appropriateness
3. Administrative Usability

CRT Technical Excellence
4. Function and Purpose
5. Objectives Development
6. Item Development
7. Methods of Score Interpretation
8. Analysis and Validation

In addition to an explanation of the rating system, this paper includes detailed instructions so that it can be used in a standardized and accurate way by school personnel, test-selection committees, researchers, and professional evaluators.

INTRODUCTION

Public education has been submitted to much scrutiny during the past decade, and the curriculum, instruction, and techniques for evaluating students and educational programs have been the focus of considerable debate. Among the subjects still being fervently discussed is the need to identify appropriate ways of measuring and describing how much and how successfully students learn in school. To many individuals concerned with testing and measurement, traditional methods do not seem to be adequate to meet this need. These people believe that although existing measures are useful for several extremely important educational purposes, like predicting who will succeed in college, they are not necessarily appropriate for many others, like describing what students have learned in school. It is within the context of an increasing demand for instructionally sensitive measures that the move toward criterion-referenced tests (CRTs) has gained momentum. Unfortunately, in their haste to develop and use CRTs, few people have paused to consider the properties of CRTs or to systematically evaluate the merit of existing ones. This paper attempts to make up for this omission by providing a system for describing and evaluating CRTs.

*This work was begun at the Center for the Study of Evaluation in the Graduate School of Education, UCLA

CRITERION-REFERENCED MEASUREMENT: A DEFINITION

A criterion-referenced test is designed to provide a measure of the extent to which an instructional task or skill has been achieved. Three of the definitions most often used are:

1. "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" (7).

2. "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed" (9).

3. "Criterion referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., a performance standard" (18).

All CRTs have several features in common:

1. They are based on clearly defined educational tasks or objectives.

2. The test items are specifically designed to measure performance on these tasks or objectives.

3. Scores are interpreted in terms of attainment of a preset criterion or level of competence with respect to the educational tasks or objectives.

THE CRITERION-REFERENCED TEST DESCRIPTION AND EVALUATION RATING SYSTEM

There are, at present, a number of tests or test systems labeled *criterion-referenced*, and these differ considerably in format, design, and function. Some CRTs, for example, consist of many small tests and are intended for classroom uses like the diagnosis and placement of students, while others contain only one or a few tests and are designed for evaluation purposes like establishing the effectiveness of an educational program. In order to provide an efficient and objective procedure for describing, assessing, and comparing all CRTs, the Criterion-Referenced Test Description and Evaluation (CRTDE) rating system was developed. A complete set of instructions to accompany the system was also prepared so that the CRTDE could be applied in a standardized way by school- and district-level personnel, test-selection committees researchers interested in CRTs for their own work, and professional evaluators.

The CRTDE system incorporates two areas of concern. First, it is concerned with the overall characteristics and usability of a CRT, including a description of a CRT's marketing and packaging features, its administrative usability, and examinee appropriateness.* Second, the CRTDE system is concerned with the specific attributes that constitute the technical quality of the test development and analysis process, including the function and purpose of a CRT, the generation of objectives and items, schemes for interpreting CRT scores, and the analysis and validation of items and tests.

It should be noted that depending upon the uses to which a CRT is put, certain items in the system will assume more or less importance. Ideally, these items would have been identified and weighted accordingly. Unfortunately, it was not possible to develop a single set of weights that would be appropriate for all CRT uses. For example, the number of test forms is always an important factor in determining the usability of a CRT, however, for classroom purposes, it is particularly desirable that the CRT consist of several short tests that can be administered throughout the year (and a high weight should be assigned to one that has this property), while for evaluation purposes, it is particularly desirable that the CRT consist of a single, comprehensive test form (and a high weight should be assigned to a CRT with this feature). In view of this, two illustrative sets of weights were developed for a subset of the CRTDE items that were thought to represent definitive properties of a CRT and, or to be crucial factors in establishing a CRT's usability. The sets of weights were patterned after two typical, but different, CRT uses. 1) as a classroom resource and 2) as a tool for evaluations involving two or more instructional programs.

In developing the CRTDE rating form, the two areas of concern – overall characteristics and technical quality – were organized into eight components that are relevant to the description and evaluation of CRTs.

Components of the CRTDE Rating Form

Overall Characteristics and Usability	1. Marketing and Packaging 2. Examinee Appropriateness 3. Administrative Usability
Technical Quality	4. Function and Purpose 5. Objectives Development 6. Item Development 7. Methods of Score Interpretation 8. Analysis and Validation

*The development of the Criterion-Referenced Test Description and Evaluation form was guided by the MEAN test evaluation procedure (12). MEAN is an acronym reflecting four critical areas of interest to test users: measurement validity, examinee appropriateness, administrative usability, and normed technical excellence. The categories of administrative usability and examinee appropriateness in the CRTDE have been particularly influenced by the MEAN procedure.

matics, in order to identify all knowledge and skills that must be acquired if the area is to be learned (11).
4. Theories of learning and instruction. A literature review is conducted and/or consultants called in to formulate series or hierarchies of educational tasks and purposes based upon the results of psychological theory and research (20).
5. Empirical studies. Experiments are conducted in order to identify the objectives that are most important because the skills and knowledge are inherently essential.

No matter how they are derived, educational tasks and purposes are usually called objectives or behavioral objectives. However, it should be noted that these terms have a precise meaning to educators: "An objective is an *intent* [author's italics] communicated by a statement describing a proposed change in a learner—a statement of what the learner is to be like when he has successfully completed a learning experience" (25). Developers of CRTs do not always use this definition in its purest sense. To them, an objective refers to the content that is supposed to have been learned (for example, equivalent and nonequivalent sets in sixth-grade math) and only sometimes includes the behaviors the student is supposed to exhibit (such as naming the first five presidents of the USA).

There are several issues involved in the formulation and generation of objectives. An important one relates to the rules needed for writing objectives and how broadly or narrowly they should be stated. Formal rules for generating and stating objectives are needed to ensure the uniformity, manageability, and comprehensiveness of the set of objectives or domain that the CRT measures.* Still another issue deals with how a domain is organized. The objectives for a single domain can be grouped by grade levels; they can be organized according to major content areas; and/or they can be arranged into a hierarchy according to the complexity of the behaviors involved or the order of instruction.

Formulating and generating items. Once the objectives for the CRT have been chosen, the next step is to construct and/or select test items to measure the objectives. This is one of the most difficult steps in the total developmental process because of the vast number of test items that might be constructed for any given objective, even those that are relatively narrowly defined (24). For example, consider the following objective: "The student can compute the correct product of two single-digit numbers greater than zero where the maximum value of this product does not exceed fifteen." The specificity of this objective is quite deceptive since there are 32 pairs of numbers that meet this requirement and at least 10 different test item types that might be used to assess student performance, as shown in Figure 1.

Further, each of the resulting 320 combinations of pairs and item types could be modified in a variety of ways that might influence whether they were answered correctly. Some of these modifications are:

- use different item formats (multiple choice versus completion)
- change the mode of presentation (written versus oral)
- change the mode of response (written versus oral)

FIGURE 1

TYPES OF CRT TEST ITEMS USING THE NUMBERS 3 AND 5

a. $\begin{array}{r} 5 \\ \times 3 \\ \hline \end{array}$

b. 5 x 3 =

c. (5) (3) =

d. 5 · 3 =

e. 5 times 3 =

f. The product of 5 and 3 =

g. 5 x __ = 15

h. If x = 5 and y = 3, what is the value of xy?

i. What number multiplied by 3 will equal 15?

j. John has 5 apples. Sally has 3 times as many apples as John. How many apples does Sally have?

It soon becomes evident that a highly specific objective can have a potential item pool of well over several thousand items (14, 15, 2).

The number of items to construct for each objective is influenced by several factors including the amount of testing time available and the cost of making an interpretation error, such as saying that a student has achieved mastery when he or she has not and therefore erroneously concluding that a student should not participate in a college preparatory program. For some objectives many items are needed in order to obtain a stable estimate of a learner's performance, whereas for other objectives fewer items will suffice.

A related issue in the construction and generation of CRT items is the degree to which the items should be sampled with respect to their relative difficulty and possible content coverage within an objective. It is a well-known and frequently used principle of test construction that even slight changes in an item can affect its difficulty. The extent to which the items within an objective are sampled with respect to difficulty has a direct bearing on the interpretation of the scores obtained. In other words, if only the most difficult items are used, the phrase "achievement of the objective" has a very different meaning than if the items are sampled over the full range of difficulties.

Another issue concerns a CRT's curriculum match—the extent to which a CRT is designed for use with a specific educational program (1, 29). CRTs with a greater degree of

*The set of objectives that a CRT measures is sometimes called a domain or universe of content (30, 5). However, the term *domain* is used by others to mean the rules for generating test items to measure a specific objective (15).

curriculum match have objectives and test items that are associated with a particular curriculum or set of educational materials and techniques. CRTs with a smaller degree of curriculum match, on the other hand, contain objectives and test items that are not necessarily associated with the specific skills or content of an educational program. However, such CRTs still may have been developed from several educational programs and consequently have objectives and items that reflect the bias inherent in these programs. Conversely, CRTs with no curriculum match are based on a domain of content and behaviors that is independent of any educational program and, therefore, can be used to compare several different educational programs.

Consideration of the various issues involved in item generation for CRTs has produced a number of different strategies for generating and constructing items. These include assembling a panel of experts (33), using content/process matrices (31); applying formal item generation rules (14, 15, 5, 27).

Formulating score interpretation schemes. The uniquely distinctive feature of a CRT is its ability to provide a means for describing what an individual (or group) can do, know, or feel without having to consider the skills, knowledge, or attitudes of others. Consequently, CRT scores are reported and interpreted in terms of the level of performance obtained with respect to the objective(s) or domain on which the CRT is based. This type of score reporting is very different from that used for norm-referenced tests in which scores are reported in terms of the performance of other individuals or groups.

1. Actual score. The number or percent of items correct on a given objective, referring to the number of items actually passed on the test.
2. True score. An individual's or group's true level of performance on an objective, referring to the portion of the total universe of items for an objective that an individual or group could answer correctly.
3. Mastery of a given objective. The achievement of a preset criterion level of performance is called mastery of an objective. Criterion levels can be selected arbitrarily or can be justified using experts' judgments and/or the results of empirical studies.
4. Performance time. The time it takes, in class hours or calendar days, for a student to achieve a given performance level.
5. Level readiness. A score that reflects the probability that the student is ready to begin the next level of instruction (this may be based on both the number of items correct and the pattern of answers given to these items).
6. Total individuals who passed. The number of individuals or groups who passed or mastered each objective or item. (This score is given most often for individual items when only one item is tested per objective—for example, National Assessment of Educational Progress.)
7. Total objectives mastered. The number of objectives passed or mastered by an individual or group.

It should be noted that scores on CRT tests need not be limited to just a CRT interpretation. Other score interpretations can also be provided to expand upon the CRT interpretation (21, 4, 8). For example, one might say that "This school had an average score of 5 out of 10 on the objective (a CRT interpretation), which is one standard deviation below the national average of 7 out of 10 (a norm-referenced interpretation). The notion of using both types of score interpretations does not reduce the theoretical soundness of the score interpretation" (4, 21, 23).

Validation of CRTs

When construction of the objectives and test items is complete, the CRT must be analyzed and validated. This process can involve giving the test to students and studying their responses (response data) or relying upon review by experts (judgmental data).

There is much ambiguity about the procedures for analyzing and validating CRTs. Nevertheless, there are several dimensions of test and item quality that are considered to be relevant to CRT validation and that have associated with them review procedures, data-collection strategies, experimental designs, and statistical indexes:

Establishing item quality. Following are several commonly considered dimensions of item quality:

1. Item-objective congruence. A test item is considered good if it measures or is congruent with the objective that it is supposed to assess. Item-objective congruence can be established by using judgmental data. Typically, content experts are given a variety of objectives and the items used to measure them and are asked to assign the items to their appropriate objective or to comment on the appropriateness of the item-objective relationship.
2. Equivalence (internal consistency within objectives). A test item is considered good if it behaves like other items measuring the same objective. The concept is similar to item-objective congruence, but its proper use depends on response data. Equivalence is usually measured by computing the biserial correlation between the score on an item and the total score on all items measuring that objective.
3. Stability (over time). An item is considered good if examinee performance is consistent from one test period to the next in the absence of any special intervention (such as instruction which is an intervention that can change examinee performance). Stability involves response data and can be measured by using a phi coefficient that correlates scores on the item from two different occasions as long as too much time does not elapse between them.
4. Sensitivity to instruction. An item is considered good if it is sensitive to instruction—that is, if there is a discrimination in responses to the item between those who have and those who have not benefited from instruction. This measure of item quality is usually computed for CRTs that are linked to particular educa-

tional programs and it requires response data. Examinees are tested before and after an educational program, and items that many examinees fail before instruction but pass after it are considered to be sensitive to the instruction.

5. Cultural/sex bias. An item is considered good if it does not lead to inaccurate conclusions about the performance of different cultural groups or sexes. Bias can be assessed using either judgmental or response data. If the former are used, representatives of different cultural groups, members of each sex, and/or linguists examine test items to determine whether vocabulary or content are foreign or could be misinterpreted. If response data are used to assess bias, they are analyzed, typically using analysis of variance or regression techniques for item-cultural/sex interactions.

Establishing test quality. There are six dimensions commonly used to express the quality of a CRT:

1. Test-objective congruence. Similar to item-objective congruence, test-objective congruence is an index of the extent to which the total test or subtest measures the stated objectives. Test-objective congruence is usually determined by using judgmental data.
2. Equivalence (internal consistency). Test equivalence is a measure of the homogeneity of test items for an objective: that is, how coherently the test items assess the particular objective. This can be measured by using split-half correlations, Kuder-Richardson formulas, or coefficient alpha.

3a. Stability (test-retest, or alternate forms). A test is stable to the extent that examinee responses are consistent from one test period to another or across alternate forms of a test in the absence of any intervention. Stability is usually measured by using correlation techniques.

3b. Stability (number of items per objective and number of objectives per domain). There are two levels at which this type of stability for a CRT can be estimated. At the first level, a determination is made of the number of items that should be tested in order to obtain a stable score on an objective. For this type of stability, the assumption is made that for each objective there is a pool or population of items with mixed difficulties that measure the objective and that for any given test a sample of those items is selected. At the second level, a determination is made of the number of objectives that should be tested in order to obtain a stable estimate of performance on the domain. For this type of stability, the assumption is made that a single score is needed that describes an individual's performance on the domain or set of objectives. Stability can be estimated with response data using correlation techniques and/or Bayesian statistics (26).

4 Sensitivity to instruction. A test's ability to discriminate between those who have and those who have not benefited from instruction. This measure of test quality is usually obtained for CRTs that are linked to a specific educational program and is obtained using response data.

5. Cultural/sex bias. Bias occurs when a test leads to inaccurate conclusions about the performance of cultural/sex groups. It can be estimated with analysis of variance or regression techniques using response data, or by expert review using judgmental data.
6. Criterion validity. Criterion validity establishes the meaningfulness of the criterion in terms of which CRT scores are interpreted. Establishing criterion validity is either a one-step or a two-step process. The first involves assessing the meaningfulness of the domain: that objectives have been selected and organized to be in themselves educationally significant and that test items have been systematically generated to cover the objectives. Step 1 criterion validity is usually established by having experts review the objectives and test items to determine the extent to which they were developed in conformance with prespecified procedures and the extent to which they cover the domain in a comprehensive and meaningful manner.

Step 1 must be completed for all CRTs, and, in some cases, is sufficient for establishing criterion validity. One example is a CRT that is based on objectives that are narrowly defined and operationally stated in such detail that generating test items only requires transposing them into question form. CRT score interpretations for objectives with these characteristics are meaningful because the objectives describe skills that can be measured directly by test items. A second example is when the CRT's objectives are linked to a curriculum and its scores are intended for and interpreted by teachers and curriculum experts. CRT score interpretations in terms of these types of objectives are meaningful because the skills and knowledge being measured are those taught in classrooms where a specific curriculum is taught.

The second step is established through empirical means and involves determining whether examinees who perform well on the test have really achieved the educational objective. Step 2 criterion validity can be measured by comparing scores obtained on a CRT by individuals who, in advance of taking the CRT and using *independent* criteria, were judged to possess or not possess the skills that the objective is intended to measure. To the extent that the CRT discriminates between these two groups of individuals, the CRT has criterion validity.*

By establishing Step 2 criterion validity, the relationship between test items and the objectives they are supposed to measure is empirically confirmed. Step 2 criterion validity permits assertions about mastery of the individual objectives that comprise a domain and about more complex behaviors whose component parts are defined by the domain. Step 2 criterion validity is particularly useful when it may be difficult to automatically assume that achievement of the items necessarily reflects achievement of the larger objective or domain.

The question of classical reliability and validity. There has been considerable debate over the appropriateness of

*Step 2 criterion validity is similar to construct validity and/or discriminant validity but an objective or a domain, rather than a psychological state, is the construct.

"classical" (long and widely used) indexes of reliability and validity to criterion-referenced tests. Some psychometricians have argued that since CRT items are selected to measure achievement of specific educational objectives and not to discriminate among students, scores on CRTs can lack variation. This could arise in the following situation: Before instruction, none of the students have mastered the objectives, and all might receive a score of zero on the criterion-referenced pretest. After instruction, however, all might receive very high scores on the criterion-referenced posttest. A lack of variation in student scores, it is claimed, would cause the traditional indexes of reliability and validity (that are based on variance) to be inappropriate (28).

Others have argued that when CRTs are administered to a heterogeneous sample representing differing degrees of competence and receiving differing instruction on the objective, there will be sufficient variation in test performance to apply the classical statistical formulas (21, 12). This latter stance is becoming the accepted view, and it is now held that the classical indexes (such as stability, equivalence) can be estimated for CRTs using a heterogeneous population.

Theoretical Value of CRTs for Evaluations

Based on theoretical considerations alone, are CRTs appropriate to measure achievement for large-scale effectiveness evaluations?

An effectiveness evaluation requires instruments that are reliable and valid and that provide meaningful scores that can be used to make decisions about educational policy. In theory, there is an orderly set of developmental and validation procedures which, if followed properly, produce CRTs that are based on well-defined sets of objectives and that can provide meaningful and useful score interpretations. Thus, from a theoretical perspective, CRTs are appropriate and desirable for measuring achievement in effectiveness evaluations.

REVIEW OF CURRENTLY AVAILABLE CRTS

Generating Review Criteria

Currently available CRTs were reviewed to determine if they were technically sound and if they could be easily used for a large-scale effectiveness evaluation. To structure the review, a set of criteria were generated. The criteria reflect the characteristics generally accepted as being necessary and appropriate for a large-scale effectiveness evaluation. Consequently, many of them could be applied to norm-referenced tests as well as CRTs. In order to obtain the criteria, several sources were consulted, including a review of the literature, requests for proposals issued by state and federal agencies involving large-scale evaluations, and criteria already developed and used for reviewing achievement tests.

Obtaining CRTs

A list of publishers of educational tests was compiled using test review books (3, 16, 17, 18), personal contacts, and library sources (24). It should be noted that publishers on the list were not necessarily known as marketers of CRTs because it was not always possible to predict in advance who published CRTs and who did not and because it was considered important to include as many publishers as possible in the review.

A letter was sent to each publisher requesting information about: any criterion-referenced math or reading tests they might have available (including detailed descriptions of the test battery at each available grade level); sample tests for reading and math at each available grade level; lists of objectives or domains for reading and math at each available grade level; directions for administering and scoring reading and math tests at each available grade level; all technical manuals, field test reports, expert reviews, or test-analysis information; information about special features like scoring services or cassette-recorded directions; and cost information.

From publishers' responses, 28 CRTs were obtained that had sufficient information for review purposes. Each CRT was independently reviewed twice using the set of criteria generated for this purpose, and discrepancies were resolved by both reviewers. Any remaining questions, usually resulting from unclear or insufficient information from the publishers, were followed up with a phone call.

Explanation of Review Criteria

There were 19 criteria against which CRTs were reviewed. (Figure 2 shows the form used by reviewers.) For this review, reading and language arts were considered to be one subject area and mathematics another. All subtests or tests of individual objectives at the same level were grouped together and considered as a single reading or math test. In addition, the criteria were especially designed to permit the cross-grade level and longitudinal comparisons that typify large-scale evaluations.

1. Coverage of specific skills. A test had to cover basic skills in reading (language arts) or mathematics.
2. Grade-level coverage. Forms of the test had to be available for grades 1 to 9 in order to make possible comparisons across grade levels as well as longitudinal comparisons. (High school-level CRTs were excluded because so few publishers had them available.)
3. Overlap of objectives across grade levels. Some or all

of the test's objectives had to be measured at each grade level in order to permit comparisons of common educational objectives across grade levels or over time.

4. Number of test forms per grade level. Due to constraints related to test administration and the time available for testing, there had to be a limited number of test forms at each grade level. Just one test per grade level was preferred in order to avoid problems with reliability that could arise when several test forms are combined.
5. Directions for test administration. A test had to provide thorough and clear instructions for both the examiner and examinee. Directions concerning distributing tests, demonstrating sample questions, and test administration had to be provided in a detailed and easy-to-read form.
6. Special equipment for test administration. Because of the logistics and costs involved in large-scale information collection, test administration could *not* involve any special equipment (like cassettes or visual aids) aside from pencils and scratch paper.
7. Time for testing. A test had to be designed to be completed within a given class period, the amount of time usually available to outside evaluators.
8. Group testing. A test had to be designed for group administration, since individual administration is prohibitive in large-scale evaluations.
9. Item-objective match. Each test item had to be coded to an objective (or the educational tasks and purposes the test claimed to measure).
10. Objective coverage. There had to be a sufficient number of items to adequately measure each objective.
11. Objective/subjective scoring. A test had to use an objective scoring procedure since it would be very costly to train individuals to use subjective scoring schemes.
12. Machine scorable. The test had to be available in, or adaptable to, machine scoring.
13. Score-interpretation scheme. A test had to employ a criterion-referenced score-interpretation scheme.
14. Reusable materials. To save money, reusable test booklets and test manuals were requested.
15. Curriculum match. A test could not be based on the objectives of any particular curriculum or educational program.
16. Costs of tests per pupil. The costs of testing pupils had to be kept low enough to accommodate a large-scale study.
17. Formal field test. A test had to provide documentation of field test activities. It was preferred that the field test participants be nationally and geographically representative, be a probability sample, and include sufficient numbers of minority persons to permit an estimation of bias.
18. Information on item quality. Information had to be provided, based either on judgmental or response data, about item stability, sensitivity to instruction, sex/cultural bias, item-objective congruence, and equivalence.
19. Information on test quality. Information had to be provided on test quality, based either on judgmental or response data, to include information about internal consistency, test stability, test-objective congruence, sex/cultural bias, sensitivity to instruction, and criterion validity.

Results of the Review

The results of the 28 tests reviewed are presented on page 8. It should be noted that because many of the 28 CRTs were intended as classroom resources and not for effectiveness-evaluation purposes, the review conducted for this investigation tended to make some CRTs look less excellent than they would have if they had been reviewed from another perspective.

1. Coverage of specific skills. Of the 28 tests reviewed, 15 were designed to assess reading skills, and 13 were designed to assess mathematics skills. All 28 tests reviewed focused on measuring basic skills in reading and/or mathematics and thus met the criterion.
2. Grade-level coverage. Nine tests were available for grades K-9, and thus met the criterion. The remainder varied from CRTs available for grades K-2 to those available for grades K-8.
3. Overlap of objectives. Twelve tests appeared to measure the same objectives at all grade levels. Sixteen tests appeared to have some overlapping objectives which were measured at most, but not all, grade levels. It should be noted that to make common objectives, test publishers frequently used broadly stated objectives or skill categories which they then translated into tasks or skills of varying complexity for different grade levels.
4. Number of test forms. Some CRTs had only one test form per grade level and others had as many as 31. Usually those CRTs that offered a limited number of test forms per grade level would include several objectives on a single test form, while those featuring more test forms per grade level would assess one or only a few objectives per form. Three tests did not set limits on the number of tests that could be created from their bank of objectives and items.
5. Directions for test administration. Twenty-seven of the tests met the criterion by providing adequate directions both to the examiner and examinee for test administration. One test contained no information about administration.
6. Special equipment required. Twenty-six tests required no special equipment for test administration and, therefore, met the criterion. Two tests required the use of tape recorders or cassettes, and one test provided no information. It should be pointed out that many of the 26 tests that do not require special equipment are designed, nevertheless, for use with special equipment and consider its omission to be undesirable.
7. Time for testing. Only two tests met this criterion. More tests (24) left time for testing open, but from their length appeared to require more than one hour of

FIGURE 2

THE REVIEW FORM

Specifications for Selecting Tests	Reviewer's Notes	Criterion Rating/Range
Coverage of specific skills Tests must cover basic skills in language arts/mathematics		P F U
Grade-level coverage Tests must be appropriate for grades 1 through 9		P F U
Overlap of objectives across grade levels Same objectives should be measured at each grade level		S A N U
Number of test forms per grade level Should be a limited number of test forms per grade level		
Complete directions for test administration		P F U
Special equipment for test administration		P F U
Time for testing Maximum of 40-50 minutes		
Group testing		P F U
Item-objective match Test items must be keyed to objectives that can be broadly stated		P F U
Objective coverage Test items should adequately cover each objective.		
Objective/subjective scoring		P F U
Machine scorable		P F U
Score interpretation Must be criterion-referenced		P F U
Reusable materials		P F U
Curriculum match Test cannot be based on specific curricula		S A N U
Cost of test per pupil		
Formal field test Preferably should have a) national scope b) geographic scope c) minority representativeness d) probability sampling		

(continued on next page)

KEY: P – pass, F – fail, S – sometimes, A – always, N – never, U – unclear, not stated, stated without supporting documentation

Specifications for Selecting Tests	Reviewer's Notes	Criterion Rating/Range
Item quality information Judgmental or response data: a) instructional sensitivity b) stability c) sex/cultural bias d) item-objective congruence		
Test quality information Judgmental or response data: a) internal consistency b) stability c) test-objective congruence d) sex/cultural bias e) instructional sensitivity f) criterion validity		

KEY: P – pass, F – fail, S – sometimes, A – always, N – never, U – unclear, not stated, stated without supporting documentation

testing time. One CRT had no information about the time needed for testing.

8. Group testing. Twenty-five tests could be administered to groups and therefore, met the criterion. Two tests were designed for individual administration only, and one did not provide this information.
9. Item-objective match. Twenty-six tests had each item coded to an objective, and one CRT did not provide this information.
10. Objective coverage. The items tested for each objective ranged from 1 to 150 across the 28 tests. (It should be noted that the CRT with 150 items per objective was based on a computerized item bank from which tests of any length could be generated.)
11. Objective/subjective scoring. Twenty-seven tests employed an objective scoring technique, meeting this criterion. One test employed a subjective technique, and one other CRT did not provide this information.
12. Machine-scoring option. Eighteen tests met the criterion for machine scoring. Nine CRTs were hand-scorable only, and one CRT did not provide this information.
13. Score-interpretation scheme. Twenty-seven tests met the criterion by using some type of criterion-referenced score interpretation scheme. Overwhelmingly, the scheme was expressed as an arbitrary mastery/nonmastery score or the number of items correct on a given objective. Of these same 27 tests, 7 also employed norm-referenced interpretations. One test did not describe its score-interpretation scheme.
14. Reusable materials. Twenty-four tests were designed so that at least some portion of the materials could be reused. These usually were the test booklets, when separate answer sheets were provided, and the teacher's and examiner's manuals. Three CRTs had no reusable materials, and one did not provide this information.
15. Curriculum match. Twenty-two tests appeared to have no match to a particular curriculum or instructional program. Six other tests also appeared to be rather general, although they claimed to be based in varying degrees on a review of what is currently being taught in today's schools.
16. Cost of tests per pupil. Based on a purchase of all tests in reading or math at the third-grade level, costs, for a minimum purchase, ranged from about five cents per student to $6.31 per student. One test had to be implemented at the district level and cost $7,500. Most tests are sold in minimum sets of 30-35 test booklets.
17. Formal field tests. Eight tests provided documentation concerning field test activities. However, the information provided was remarkably sparse with several exceptions. Those who did conduct field tests usually attempted to get some sort of geographic and national representation. Fifteen tests claimed to have been field tested, but provided no supporting documentation. Five additional tests provided no information at all about field tests.
18. Information on item quality. Twelve test publishers reported having conducted item quality studies based on response data and/or expert review. Of these, attention typically was paid to item-objective congruence, item stability or equivalence, and sensitivity to instruction. Eight tests reported having some type of review but declined to state the kinds or extent of their studies. Eight other systems did not provide any information at all.
19. Information on test quality. Thirteen tests reported having conducted test-quality studies based on response data and/or expert review. Of these, internal consistency, stability, test-objective congruence, sensitivity to instruction, and criterion validity (Step 1) were most frequently attended to. Seven other systems claimed to have performed test quality studies, but provided no supporting documentation. Eight additional systems provided no information at all.

Practical Value of CRTs for Evaluations

Based on practical considerations alone, are CRTs appropriate for large-scale effectiveness evaluations?

The answer is no. From the review, it is clear that no CRT fully met all the criteria. Further, the review uncovered a number of serious practical problems that diminish the suitability of currently available CRTs for an effectiveness evaluation:

Many learning objectives. Most of the CRTs reviewed had a large number of very specific learning objectives that were associated with very small units of instruction, like one to five class lessons. The reason for the use of many narrowly defined objectives can probably be traced to the original use of CRTs by teachers as an aid to individualizing and evaluating instruction. Nevertheless, an effectiveness evaluation of the impact of just one year of instruction at one grade level would generate information about an enormous number of objectives, thus complicating the management, analysis, and reporting of data.

Numerous test forms. Many currently available CRTs provide separate test forms for each grade level that measure just one or a few different objectives. The appearance of many test forms also probably reflects the original intention to use CRTs as classroom aids. In terms of an effectiveness evaluation, the logistics of administering a number of distinct tests complicates information-collection activities and increases the chances of making errors and the costs of conducting the evaluation.

Maximum time required for testing. Most available CRTs take more than an hour of class time, which is the maximum time that can usually be devoted. It should be noted that some of the test publishers, recognizing time constraints, offered CRTs that had just one item per objective. However, this is not a satisfactory solution since reducing the number of items will almost invariably bring with it a diminution in the test's ability to measure with precision each of the objectives although it may have the beneficial effect of diminishing testing time.

Discrepancies between CRT's and program's objectives Using CRTs in effectiveness evaluations that involve more

TABLE 1

TESTS REVIEWED

Name of System	*Publisher*
Fountain Valley Teacher Support System-Reading	Richard Zweig Associates, Inc.
Fountain Valley Teacher Support System-Mathematics	Richard Zweig Associates, Inc.
Prescriptive Reading Inventory	CTB/McGraw-Hill
Diagnostic Mathematics Inventory	CTB/McGraw-Hill
Comprehensive Tests of Basic Skills Form S (CTBS/S)-Reading	CTB/McGraw-Hill
Comprehensive Tests of Basic Skills Form S (CTBS/S)-Mathematics	CTB/McGraw-Hill
ORBIT (Objective's-Referenced Bank of Items and Tests)	CTB/McGraw-Hill
Skills Monitoring System-Reading*	Harcourt, Brace, Jovanovich, Inc.
1973 Stanford Reading Tests	Harcourt, Brace, Jovanovich, Inc.
1973 Stanford Mathematics Tests	Harcourt, Brace, Jovanovich, Inc.
Individualized Criterion-Referenced Testing-Reading	Educational Development Corp.
Individualized Criterion-Referenced Testing-Mathematics	Educational Development Corp.
Woodstock Reading Mastery Tests—Form A	American Guidance Service
Key Math (Diagnostic Arithmetic Test)	American Guidance Service
Mastery: An Evaluation Tool, SOBAR, Reading	Science Research Associates
Mastery: An Evaluation Tool-Mathematics	Science Research Associates
Individual Pupil Monitoring Systems-Reading	Houghton-Mifflin
Individual Pupil Monitoring Systems-Mathematics	Houghton-Mifflin
Comprehensive Achievement Monitoring (CAM) Maintenance Pkg.-Reading	National Evaluation Systems
Comprehensive Achievement Monitoring (CAM) Maintenance Pkg.-Mathematics	National Evaluation Systems
Objectives-Based Test Sets-Reading	Instructional Objectives Exchange
Objectives-Based Test Sets-Mathematics	Instructional Objectives Exchange
Reading-Analysis of Skills	Scholastic Testing Service
Mathematics-Analysis of Skills	Scholastic Testing Service
Tests of Achievement in Basic Skills (TABS)-Reading	Educational and Industrial Testing Service
Tests of Achievement in Basic Skills (TABS)-Mathematics	Educational and Industrial Testing Service
Reading Inventory Probe I	American Testing Company
Mathematics Inventory Tests	American Testing Company

*This test was not available at press time.

than one educational program means determining relationships between the general objectives the CRTs are designed to measure and those of the programs so that achievement can be measured in terms of the objectives emphasized in instruction and exemplary programs can be identified. However, obtaining this information is costly and complicated. Teachers can be asked, for example, to rate the CRTs' objectives in terms of their relevance to classroom instruction, but teacher ratings can be unreliable. Instructional experts can be asked to analyze textbooks and curriculum guides, however, they cannot know for certain how these materials are being used in the classroom.

A related problem concerns which objectives to test. Each student or classroom can be tested on just those objectives derived from the curriculum being used or on a sample of objectives some of which may be relevant to the curriculum while the others are not. Depending upon the choice, the resulting evaluation information can be limited in its usefulness for making comparisons or it may require considerable manipulation before interpretations can be made.

Identifying common objectives across grade levels. The same objectives are not always measured at all grade levels or, if they are, there is no system for identifying common objectives. The skills and content associated with an objective generally become more complex at higher grade levels. To make comparisons over time or across grades, however, it is necessary to identify skills or objectives that are related in terms of a conceptual framework or general content area. For example, in the fourth grade, a punctuation objective might focus on beginning sentences with capital letters and ending them with periods, while in the ninth grade, a punctuation objective might focus on the proper use of semicolons as alternatives to periods. Unless the test publisher has identified the relationship between these two objectives—for example, that they both have to do with the same skill area—the evaluator may be forced to decide this on his own, an instructional decision that is not ordinarily part of the evaluator's expertise.

Unvalidated CRTs. The procedures used to validate CRTs are not very sophisticated, and field test results are not reported in any detail. When compared with the highly structured field tests conducted for norm-referenced tests, most CRTs are deficient with respect to the sample's size and representativeness, and/or the amount and precision of data presented in technical reports.

Insufficient score information. Most CRTs report scores either as the number of items correctly answered for each objective or sometimes as mastery or nonmastery scores, "mastery" meaning correctly answering an arbitrarily selected number of items per objective. These types of score interpretations are accepted by theorists as legitimate ways of expressing CRT test scores and they may have meaning for teachers who know their curriculum. However, for effectiveness-evaluation purposes, these types of interpretations are inadequate because they provide insufficient information for decision making and lose meaning outside the classroom.

Financial considerations. A final practical problem with using currently available CRTs for effectiveness-evaluation purposes is that most are costly. This probably reflects the effort it takes to define domains, to develop the special features offered by CRTs, such as referencing the objectives to various school curriculums, and to provide many short test forms that can be used efficiently for classroom instruction purposes.

CONCLUSIONS

There is no currently available CRT that is feasible for use in large-scale effectiveness evaluations. This conclusion is based on practical, not theoretical, considerations. One major reason for the likely inappropriateness of available CRTs is that many of them have been designed for classroom and not evaluation purposes, consequently, they are characterized by numerous, narrowly defined objectives, each measured on a separate test form. In the context of an effectiveness evaluation, these CRTs produce unwieldy amounts of information, require too much time for testing, and create logistical problems for test administrators.

A second major practical failing of currently available CRTs is that field tests are either not documented or are performed inadequately. As a result, the reliability and validity of these CRTs is simply not known, and it is inappropriate to provide decision makers with information of unconfirmed quality.

A third major failing of available CRTs is that the score interpretations given are not as meaningful as can be expected. Most are presented as numbers of items passed, without Step 2 criterion validity information or comparative data as supplements. Two additional practical failings are the CRTs' costs and the absence of mechanisms for tracking the same skills or objectives across grade levels.

A CRT that is appropriate to use in measuring achievement in an effectiveness evaluation should be based on a limited set of objectives that represent essential competencies and basic skills, should be proven reliable and valid, and should be able to provide scores that are meaningful and useful.

REFERENCES*

1. Baker, R.L. Measurement considerations in instructional product development. Paper presented at Conference on Problems in Objective Based Measurement, Center for the Study of Evaluation, UCLA Graduate School of Education, Univer. of California, Los Angeles, 1972.

2. Bormuth, J.P. *On the theory of achievement test items.* Chicago: Univer. of Chicago Press, 1970.

3. Buros, O.K. (Ed.) *The mental measurements yearbook.* Highland Park, N.J.: Harper, 1970.

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

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

6. Cronbach, L.J., & Suppes, P. (Eds.) *Disciplined inquiry for education.* Stanford, Calif.: National Academy of Education, 1969.

7. Davis, F.B., & Diamond, J.J. The preparation of criterion-referenced tests. *CSE Monograph No. 3.* Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, Univer. of California, Los Angeles, 1974.

8. Ebel, R.L. Evaluation and educational objectives: behavioral and otherwise. Paper presented at the Convention of the American Psychology Association, Honolulu, Hawaii, 1972.

9. Fink, A., & Kosecoff, J. *An evaluation primer.* Book in preparation, 1976.

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

11. Glaser, R., & Nitko, A. Measurement in learning and instruction. In R.L. Thorndike (Ed.). *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971. Pp. 652-670.

12. Harris, C. Comments on problems of objective-based measurement. Paper presented at the annual AERA meeting, New Orleans, 1973.

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

14. Hively, W. Introduction to domain referenced achievement testing. Symposium presentation. AERA, Minnesota, 1970.

15. Hively, W., Maxwell, G., Rabehl, G., Senion, D., & Lundin, S. Domain referenced curriculum evaluation: A technical handbook and case study from the MINNEMAST project. *CSE Monograph No. 1.* Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, Univer. of California, Los Angeles, 1973.

16. Hoepfner, R. *et al. CSE Elementary school test evaluations,* Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, Univer. of California, Los Angeles, 1971.

17. Hoepfner, R. *CSE Secondary school test evaluations: Grades 7 & 8.* Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, Univer. of California, Los Angeles, 1974.

18. Hoepfner, R. *et al. CSE-ECRC Preschool/Kindergarten test evaluations.* Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, Univer. of California, Los Angeles, 1971.

19. Hofstadter, R. *Anti-intellectualism in American life.* New York: Vantage Books, 1963.

20. Keesling, J.W. Identification of differing intended outcomes and their implications for evaluation. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.: 1975.

21. Klein, S.P. Evaluating tests in terms of the information they provide. *Evaluation Comment,* 1970, 2(2), 1-6 ED 045 699.

22. Klein, S.P. An evaluation of New Mexico's educational priorities. Paper presented at Western Psychological Association, Portland, 1972. TM 002 735 (ED number not yet available).

23. Klein, S.P., Fenstermacher, G., & Alkin, M. The center's changing evaluation model. *Evaluation Comment,* 1971, 2(4).

24. Klein, S.P., & Dosecoff, J.B. Issues and procedures in the development of criterion-referenced tests. ERIC/TM Report 26. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1973.

25. Mager, R.F. *Preparing instructional objectives.* San Francisco: Fearon, 1962.

26. Novick, M.R., & Lewis, C. Prescribing test length for criterion-referenced measurement. *CSE Monograph No. 3.* Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, Univer. of California, Los Angeles, 1974.

27. Popham, W.J. *Educational evaluation.* Englewood Cliffs, N.J.: Prentice-Hall, 1975.

*Items followed by an ED number (for example ED 045 699) are available from the ERIC Document Reproduction Service (EDRS). Consult the most recent issue of *Resources in Education* for the address and ordering information.

28. Popham, W.J., & Husek, T.R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, 6(1), 1-9.

29. Skager, R. Generating criterion-referenced tests from objectives based assessment systems: Unsolved problems in test development, assembly and interpretation. Paper presented at the annual AERA meeting, New Orleans, 1973.

30. Skager, R. Critical differentiating characteristics for tests of educational achievement. Paper presented at the annual AERA meeting, Washington, D.C., 1976.

31. Wilson, H.A. A humanistic approach to criterion-referenced testing. Paper presented at the annual AERA meeting, New Orleans, 1973.

32. Wilson, H.A. A judgmental approach to criterion-referenced testing. *CSE Monograph No. 3*. Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, University of California, Los Angeles, 1974.

33. Zweig, R., & Associates. Personal communication, March 15, 1973.