An NCME Instructional Module on

Using CRTS in Program Curriculum Evaluation

Steven J. Osterlind
University of Missouri–Columbia

This monograph describes how criterion-referenced tests (CRTs) can be used in program and curriculum evaluations for developing information to form judgments about educational programs and curricula. The material is organized as follows: a brief introduction to the monograph, its purpose and goals, a discussion of the relationship between evaluation and CRTs, pertinent information about program and curriculum evaluation, relevant facts about CRTs, and a summation. A principal goal of the essay is to describe concepts and procedures in terms that are instructionally illuminating. The reader is guided to identify and examine particular points at which decisions must be made about how, when, and why CRTs may aid the evaluation process. These steps are each identified as "An Instructional Step" and presented in separate boxes with pertinent guiding questions. Conciseness is a further aim of this monograph, one consequence of which is that several important concepts are only cursorily described or alluded to. Annotated references are included. Accompanying this instructional monograph is a "Student's Self-Test." *An* "Instructor's Guide" *with expanded references and materials for photocopying or preparing transparencies is available by mail order (see "Teaching Aids" ordering information).*

Steven J. Osterlind is Associate Professor and Director, Center for Educational Assessment, University of Missouri–Columbia, 403 South Sixth St., Columbia, MO 65211. He specializes in measurement.

Series Information

ITEMS is a series of units designed to facilitate instruction in educational measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes. Barbara S. Plake, University of Nebraska–Lincoln, has served as the editor for this module.

In educational settings, evaluation is the process of determining the worth of an instructional program. Such evaluative judgments typically require information about particular outcomes or effects, or the attainment of objectives. One evaluation tool that is especially helpful in this regard is the criterion-referenced test (CRT), a measurement instrument constructed to yield information about specified performance standards. The scores on a CRT may be used as a criterion measure for making particular evaluative judgments about programs and curricula. Of course, complete judgments for evaluation will require information beyond just that which may be inferred from scores on a CRT. Nevertheless, CRTs, when properly understood, appropriately installed, and correctly interpreted, can contribute significantly to many program and curriculum evaluations.

This monograph describes how CRTs can aid the process of making evaluative judgments about educational programs and curricula. It is important to realize from the outset that evaluation is a broadly based process encompassing a number of methodologies, each of which may employ a variety of tools; the CRT is but one of these aids. The enormous importance of CRTs in program and curriculum evaluation, however, warrants special attention.

A principal goal of this essay is to describe concepts and procedures for using CRTs in program and curriculum evaluation in terms that are instructionally illuminating. To this end the reader is guided to identify and examine particular points at which decisions must be made about how, when, and why CRTs may aid the evaluation process. These points are each identified as "An Instructional Step" and are presented in separate boxes with pertinent guiding questions. A number of checklists, tables, and charts are also included to serve as quick reference to specific instructional items.

Conciseness is a further aim of this monograph. One consequence of brevity is that several important concepts are described only cursorily or are alluded to in passing. Because thoroughness of coverage is not possible in this brief introduction to the topic, a carefully selected list of references important to a more complete understanding of how CRTs may be used in program and curriculum evaluation appears at the end of the monograph.

Instruction

The Relationship Between Evaluation and CRTs

Evaluation is a process of determining the worth of programs and curricula; it requires a series of activities to be systematically conducted according to principles commonly agreed upon by people engaged in the profession. In practice, evaluation produces information of a particular type depending on the evaluation design used. One kind of evaluation approach will yield data about student performances, whereas another will produce information about teaching strategies; yet another design may be aimed at providing helpful information when program goals are unclear or are unrealistic. There are many frameworks, or *models*, for evaluation of programs and curricula, each targeted at producing a certain kind of information.

Most evaluation designs require measurement instruments, such as questionnaires, interviews, surveys, observations, or a variety of tests. Standardized tests of achievement, ability, or aptitude are probably the most frequently used of the quantitative evaluation tools available. Until recently, evaluations that used a standardized test routinely employed norm-group referenced tests (NRT). Within the last 10 or more years, however, there has been a dramatic increase in the use of CRTs in program and curriculum evaluation.

The increased use of CRTs in program and curriculum evaluation is probably due to the fact that CRTs are designed to yield data targeted at a narrowly defined, often idiosyncratic, question. Such high-focus capability of CRTs makes them ideally suited for program and curriculum evaluation, because most evaluations seek to make judgments about specific programs or curricula. For example, a school district administrator may wish to learn about the effects of just one reading program—the one used in his or her school district—and would be less interested (for purposes of the evaluation) in knowing about generalities among various reading programs. CRTs are one powerful tool that can be used in evaluations of this type.

Understanding Program and Curriculum Environments

There are many and diverse considerations in planning for an evaluation. Some of these considerations may be choosing just one program to be evaluated from among the possible programs that could be evaluated, considering the right objectives for evaluation, determining how one can use the available resources in an evaluation, understanding how program components interact to produce results, and examining outcomes. Of course, this is only a small sample of the potential number of things to consider when decisions about an evaluation are made. Each consideration requires careful deliberation and has significant implications for selecting an appropriate evaluation design. Often, the choices made in planning for an evaluation will guide next steps about what tools should be used, that is, whether to use a CRT or something else. One way to begin the process of asking the right questions is to understand the context in which a program that is to be evaluated resides.

Contexts within which educational programs or curricula operate may be referred to as *program environments*. Knowing the program environment will influence the kinds of questions that can be addressed in the evaluation, as well as the methodology that can be used. In particular, three kinds of program environments exist for most program and curriculum evaluations: regular school district programs, special compensatory programs, and other discretionary programs.

Regular school district programs are those in which nearly everyone enrolled at a school participates; they are what the public at large imagines to be school experiences, such as the grade 6 reading program. These programs and curricula are distinctly local in focus, although they may be very similar for many school districts. Extreme programs are rare. Their development is idiosyncratic in that a given school district may dictate programs and curricula by administrative mandate, whereas another school could have broad participation by administrators, teachers, parents, and even students in deciding the curriculum, its approach, and its emphasis.

Until recently, NRTs were used almost exclusively in program and curriculum evaluations of regular school district programs; however, the use of CRTs in evaluations of these programs has increased enormously. This trend suggests that many people want the kinds of information that will address individual questions. Because CRTs can accommodate individual needs for information better than the defined group-referenced interpretation of scores on an NRT, CRTs may be more appropriate for instrumentation in most evaluation efforts of programs and curricula of regular school district programs.

A second type of program environment is special compensatory programs, such as the Education Consolidation Improvement Act of 1981 (ECIA) Chapter 1, Title VII, and Title IX. Bilingual education programs and services for the handicapped are also special compensatory programs. In most instances, these programs are directed at the state or federal level, and the evaluation design and process is dictated to local districts. For example, Title I Evaluation Reporting System (TIERS) is a federally mandated program evaluation to which local school districts must conform. Researchers report that evaluation activities for special compensatory programs typically have little relevance to local audiences and are often more directed at providing information to the funding agency.

NRTs, with their emphasis upon the possible comparisons to a standardization group that is nationally representative (a focus of attention for many funding agencies), are typically

Teaching Aids Are Available

A set of teaching aids, designed by Steven J. Osterlind to complement his ITEMS module, "Using CRTs in Program Curriculum Evaluation," is available at cost from NCME. These teaching aids consist of an "Instructor's Guide" with expanded references and copies of the figures and tables used in this module in various formats suitable for photocopying and preparing transparencies. As long as they are available, they can be obtained by sending $5.00 to: **Teaching Aids, ITEMS Module #5, NCME, 1230 17th St., NW, Washington, DC 20036.**

used in evaluations of special compensatory programs. Nevertheless, as federal and state regulations and guidelines grow more flexible, more and more local school districts are using CRTs in evaluation of special compensatory programs.

A third kind of program environment is other discretionary programs. These programs may have their impetus from a state or even a federal agency, but more typically they are devised and funded locally. They may be experimental programs, demonstration programs, or other types of optional programs. Rarely is the funding for these programs assured for a long term, and some researchers conclude that the success of many of these programs is due to the efforts of one or two charismatic, dedicated persons. Because these programs are often idiosyncratic and unique, it is a natural fit for CRTs to become a predominant part of program and curriculum evaluations of discretionary programs (see Figure 1).

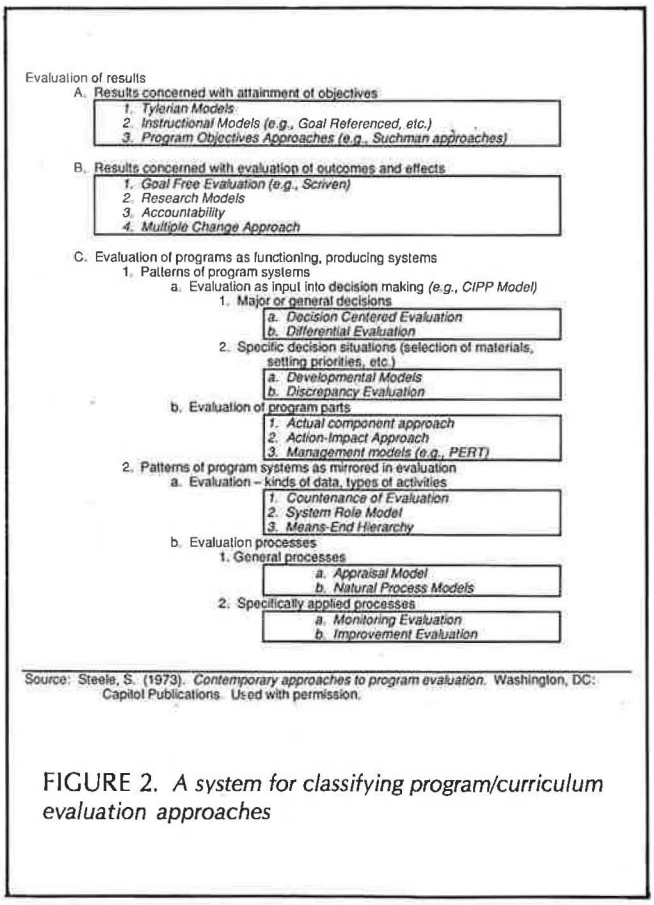
Determining an Appropriate Evaluation Model

There is, as one can imagine, a wide array of evaluation models available to assist one in organizing a search for answers to questions that will lead to determining the worth of a program or curriculum. Many, but not all, evaluation approaches allow for using a CRT. Deciding which evaluation models do and which do not can be a tricky issue. For example, when the concern for a given program or curriculum evaluation is to set priorities for distributing limited resources, the appropriate evaluation model and the proper use of a CRT is different than, say, when one is determining the right level of objectives for a given instructional program, or different again when the issue is to find out why a program is not producing as much as was expected, or different still when the question of interest is to improve everyday operations of a program.

Asking clearly thought-out questions before determining which evaluation tool to use is an important step. One simply cannot conduct an evaluation using a haphazardly chosen model or an incorrectly installed CRT and then expect that any question at all may be answered by the information produced.

Happily, the various approaches to contemporary program and curriculum evaluation may be categorized by criteria pertinent to using a CRT in the model. One such taxonomy uses six criteria to guide its schematization, including (a) evaluation that will help make major decisions and guide overall program management, (b) evaluation for examining the impact and larger results of programs, (c) evaluation to guide the organization and use of program components, (d) evaluation for viewing the program through the eyes of the participant, (e) evaluation for examining the results of instruction, and (f) evaluation that can serve other specific purposes.

These criteria yield a system for classifying evaluation approaches that will help one understand and appreciate the phenomena being classified. The taxonomy is presented in Figure 2 (which includes only the portions of the taxonomy relevant to evaluating educational programs and curricula). Also presented in Figure 2 (in boxes) is a suggested evaluation model for a particular purpose. By examining the table, one may begin to get a feel for the relationship between evaluative questions and evaluation models. It is emphasized that



Source: Steele, S. (1973). *Contemporary approaches to program evaluation.* Washington, DC: Capitol Publications. Used with permission.

FIGURE 2. *A system for classifying program/curriculum evaluation approaches*

FIGURE 1. *Instructional step number 1*

Determine in which program environment you may install a CRT for program or curriculum evaluation by asking these questions:

- Is this program a regular district program?
- Is this program a special compensatory program?
- Is this program a discretionary program?

the suggested models are not a complete listing of all the possible models. For each level in the taxonomy, there are many alternative approaches to evaluation (see Figure 3).

Understanding What CRTs Are

A criterion-referenced test is one that is deliberately constructed to yield measurements directly interpretable in terms of specified performance standards. Three requisite features for CRTs are (a) clearly defined performance standards for measurement, (b) test items constructed specifically

Determine from the above classifying criteria and from Table 1 a suggested approach to evaluation by following these steps:

- Review the criteria for classifying evaluation approaches as a prompt to asking yourself the relevant questions for your particular evaluation.
- Find the portion of the taxonomy that most nearly addresses your evaluation concerns.
- Read in the taxonomy a suggested approach to evaluation. [NOTE: if you are unfamiliar with the evaluation tool suggested, you may find a description of the model in one or more of the references cited in the References section of this monograph.

FIGURE 3. *Instructional step number 2*

to address the intended performance standards, and (c) scores that can be interpreted in terms of an individual's achievement of the specified performance standards.

Often, a CRT is viewed in perspective by delineating characteristics opposite an NRT. The distinction between a CRT and an NRT is acute. In an NRT, performance is gauged by comparison of a test score to that of the reference group. For example, on a test of reading ability, the NRT yields information about how one performs a reading activity *compared with* how well others perform the same activity. Frequently, the domain for measurement is not articulated with precision, and sometimes it is not even stated at all—the description of an NRT as one of "reading achievement" may be about as precise as is offered. Scores are reported as comparative, derived scores, such as percentile ranks, stanines, and grade equivalents.

In CRTs, on the other hand, the domain "reading ability" is articulated into observable performance, such as carrying out written instructions, rephrasing sentences, or reacting emotionally to described events. The domain can be specified in elaborate, even laborious, detail. The point of CRTs is to provide information about how well an individual can perform such tasks, not how much comparative "reading ability" the individual may possess along a hypothetical ability dimension (see Figure 4).

There are two general categories of CRTs. CRTs may be classified as having well-defined and *ordered* domains, or they may possess well-defined but *unordered* domains. A domain is well defined if it is clear to knowledgeable persons which kinds of tasks may be considered for potential test items. Domains may be ordered or unordered based on judgments of the social or aesthetic quality of an examinee's product or performance. An ordered domain is one that is amenable to being scaled by a numbering system, implying nuances of quality. For example, a scale can be assigned to a test of handwriting quality, with criterion levels established ranging from poor to exceptional legibility. The domain of handwriting is thus "ordered."

Unordered domains, on the other hand, although equally acceptable for CRTs, are representations of learning outcomes that cannot be ordered, for example, the latent trait of reading. Although particular reading tasks can be identified and articulated, there is not a strictly linear relationship among tasks wherein each step is prerequisite to the next. An examinee may be able to identify metaphor accurately without being able to distinguish between fact and opinion, for example. Metaphor and fact versus opinion are simply two of many component parts to complex reading, and neither is a prerequisite to learning the other. Most of the literature on CRTs deals with tests based on unordered domains like reading. The key is that they are well-defined domains. As has been emphasized, a well-defined domain is a necessary condition for criterion referencing.

A scheme for classifying and distinguishing criterion-referenced tests based on clarity of domain specifications is presented in Table 1. Note in the headings of Table 1 that there are two additional columns not yet mentioned: ill-defined domains and undefined domains. There are many tests available, both commercially developed and more locally devised, that fall into these categories. Despite claims by some of the developers to the contrary, tests distinguished by these classifications cannot be considered CRTs (see Figure 5).

Standards for Determining the Quality of a Particular CRT

As one may expect, CRTs—those commercially available as well as those locally developed—vary dramatically in technical merit. Many excellently constructed CRTs are available, and many more are built by competent persons for particular purposes; one should be cautioned, however, that there are many more poorly constructed CRTs. Further, it is difficult to determine which test is well done and which is not. A test instrument cannot be evaluated by a superficial inspection of a finished test booklet. A more thorough investigation into the procedures used for construction is required. Fortunately, there are widely accepted industry standards for judging educational and psychological tests. These are the *Standards for Educational and Psychological Testing* (American

FIGURE 4. *Instructional step number 3*

Determine whether the measure you are considering is criterion-referenced by asking these questions:

- What is the primary purpose for which this test was constructed
- How was it constructed?
- What is the specificity of the information yielded about the domain of instructionally relevant tasks?
- What is the generalizability of test performance information to the domain? [NOTE: this question is applicable to most, but not all, CRTs.]
- What use will be made of the obtained test information?

TABLE 1

A Scheme for Classifying and Distinguishing Criterion-Referenced Tests

Basis for Test Development	Well-defined and Ordered Domains	Well-defined But Unordered Domains	Ill-defined Domains	Undefined Domains
	Ordering based on judgements of the social or aesthetic quality of an examinee's product or performance	Specifying the stimulus properties of the items to be included in the domain	Poorly articulated behavioral objectives	No attempt to define a domain to which test performance is referenced
	Ordering based on which level of difficulty or complexity a topic or subject is learned	Specifying the stimuli and the responses in the domain	Defining the domain only in terms of the particular items on the test	Using a cut-off score, but not defining a performance domain
	Ordering based on degree of proficiency with which a complex skill is performed	Specifying the "diagnostic" categories of the domain		
	Ordering based on prerequisite sequences for acquiring an intellectual or psychomotor skill	Specifying the abstractions, traits, or constructs that define the domain		
	Basis for Test Development			
	Ordering based on an empirically defined latest test	Other ways of specifying the domain are possible		
	Ordering on other bases is possible			

Source: Nitko, A. J. (1980). A scheme for classifying and distinguishing criterion-referenced tests. *Review of Educational Research, 50,* 461-485. Used with permission.

Educational Research Association, American Psychological Association, & NCME, 1985). The *Standards* are comprehensive in that they address the three main participants in the testing process: the test developer, the test user, and the test taker. The *Standards* are not intended to prescribe precise statistical methods; rather, they provide criteria for the development of test instruments, their use, and their interpretation.

In addition to using the *Standards* as a guide to judge the quality of a test, there are other aids. One aid that will make the task of judging a test more manageable is an outline of relevant questions. One suggested outline for evaluating a possible instrument is presented in Figure 6. Identifying and studying the test characteristics called for in the categories outlined in Figure 6 is one practical way to gauge the quality of any CRT that may be considered for installation in a program or curriculum evaluation.

FIGURE 5. *Instructional step number 4*

Determine classificatory features of the CRT you are considering by asking these questions:

- Are the domains to be measured well-defined, ill-defined, or undefined? [Note: use the "Basis for Test Development" criteria listed in Table 2 to guide you.] If the domains are ill-defined or undefined, the measure is not a true CRT and is inappropriate for the present purposes.
- Is the domain ordered? If so, then find the basis for scaling or ordering the defined domain.
- Is the domain unordered? If so, then find the basis for delineating the behavior domain and the area for emphasis during test development.

Recognizing the Role for CRTs in Common Evaluation Models

There are, of course, literally thousands of CRTs (of varying quality) available and hundreds of evaluation models (of varying rigor and technical merit) eligible for evaluating a given program or curriculum. The possible combinations of matching CRTs to evaluation models that could be described is nearly infinite. Of course, not just any CRT can be used in any evaluation model. Certain models have particular requirements or restrictions that will allow or prohibit the use of a given CRT.

One family of evaluation approaches that relies heavily on CRTs is so popularly used that it deserves specific mention. This family of evaluation approaches is the Tylerian Models for Evaluation, with their orientation to objectives, testing, and experimental design. There are many evaluation designs that fall into the family of Tylerian models for evaluation. It is beyond the scope of this monograph to describe operational aspects for each; instead, references that include fuller descriptions of these models (as well as many others) are cited in the References section at the end of the monograph (e.g., Berk, 1980, 1984). The reader is invited to explore particular models in these citations. The point to remember here is that CRTs are especially well suited to evaluation models that rely on objectives, testing, and experimental design.

In addition to the paramount importance of applying a CRT for instrumentation in Tylerian models, there remain other popular approaches to evaluation that also allow for install-

A. *General Information*
- Title of instrument (including edition and forms if applicable)
- Author(s) and publisher, date of publication
- Time required to administer
- Cost

B. *Brief Description of Purpose and Nature of the Instruments*
- General type, nature of content
- Population for which designed
- Subtests and separate scores, types of items

C. *Practical Evaluation*
- Qualitative features (design, ease of use, attractiveness, etc.)
- Ease of administration, clarity of directions
- Scoring procedures
- Administrator qualifications and training
- Face validity and examinee rapport

D. *Technical Evaluation*
1. *Norms*
   - Type of norm-based scores
   - Standardization sample (size, representativeness)
2. *Reliability*
   - Type and procedure including size and nature of samples employed
   - Scorer reliability if applicable
   - Equivalence of forms
   - Long-term stability
3. *Validity*
   - Specification of variable supposed to be measured
   - Appropriate types of validation procedures (content, criterion-related, construct)
   - Specific procedures followed in assessing validity and results obtained
   - Size and nature of samples employed

E. *Reviewers' Comments*
- From *Mental Measurements Yearbooks*, journal reviews, or other sources

F. *Summary Evaluation*
- Major strengths and weaknesses across all categories

Source: Anastasi, A. (1982). *Psychological Testing* (5th ed.). New York: MacMillan. Used with permission.

FIGURE 6. *A suggested outline for test evaluation*

ing a CRT. Some of these are Decision-Centered Evaluation models (as, for example, CIPP), Countenance of Evaluation, Discrepancy Evaluation, and Goal-Free Evaluation. All these approaches to evaluation allow for installing a CRT. Finally, it should be noted that there are other important evaluation models that are not addressed here because the methodology does not allow for installation of a CRT. Two examples of these evaluation models are the Adversary or Judicial Evaluation Model and the approaches to evaluation described by Elliot Eisner as Educational Connoisseurship and Criticism (see Madaus, Scriven, & Stufflebeam, 1983).

Unquestionably, the mainstay of educational evaluation theory for more than 40 years has been the Tylerian Evaluation Rationale. The influence of Tyler on the evaluation scene has been enormous, ranging from the now-famous Carnegie Eight-Year Study (1932–40) to the initial conception of the National Assessment of Educational Progress (NAEP). In the Tylerian approach, evaluation is concerned with making a determination of whether student attainments matched one or more defined instructional or behavioral objectives. The essential role for evaluation is to improve programs and curriculum by exploring how far the instructional activities go toward actually achieving the desired results. The process for evaluation includes analyzing objectives to identify their behavioral content, identifying situations or circumstances in which a student could be expected to exhibit those behaviors, and selecting or developing test instruments from which it may be inferred whether the behaviors took place. Clearly, the major ingredient for evaluation is a standardized test. Until the advent of CRTs in the mid-1960s, the tests used with Tylerian approaches to evaluation were almost exclusively of the normative-referenced kind. In most instances today, however, CRTs are more apt for installation into a Tylerian-based evaluation of educational programs or curricula than are norm-referenced measures. The reasons for this are threefold: (a) CRTs require a higher degree of specificity for objectives than do NRTs, (b) they are deliberately constructed to assess particular objectives, and (c) they yield information directly interpretable in terms of the objectives. Hence, each of the features of CRTs are the essence of Tylerian evaluation models.

The notion of specifying objectives as the principal feature in an instructional effort has been carried further in recent years. W. James Popham (1984), of the University of California, Los Angeles, and director of the Instructional Objectives Exchange (IOX), a long-time advocate of increased specification for the intent of a given instructional effort, argues that behavioral objectives, regardless of clarity, are insufficient to define a domain unambiguously because they permit too much variation in their content, form, and difficulty. Popham (1984) posits the necessity for focusing on diagnostic categories that may be articulated as *test content specifications*. Test content specifications are elaborated statements of performance standards detailing stimulus and response attributes for particular test items. When these test content specifications are evidenced in a properly conceived CRT and installed in a program or curriculum evaluation, they will de facto lead instruction in a prescriptive manner. Popham labels such a scheme "measurement driven instruction" and claims that "if properly conceived and implemented, measurement driven instruction currently constitutes the most cost-effective way of improving the quality of public education in the United States" (1987, p. 679). Measurement driven instruction is extant only because of CRTs.

It is obvious from a casual inspection that CRTs are the quiddity of objectives-based program evaluation models. The implementation of these models, whether by Tylerian frameworks for evaluation or by other instructional strategies, such as measurement driven instruction, requires a CRT. Answers to the specific questions sought by particular objectives-based evaluation models are directly supplied by CRT-type information. CRTs, when understood and properly used in a program or curriculum evaluation, are a "can-do" strategy to achieve the desired judgmental aim for objectives based evaluation (see Figure 7).

Standards for Determining the Quality of a Particular Evaluation

It was noted earlier in this monograph that there are accepted standards for judging the technical merits of test instruments and addressing concerns about their construction, use, and interpretation. There are also standards available for judging the merits of educational programs and curricula. These are the *Standards for Evaluations of Educational Programs, Projects, and Materials* (Joint Committee, 1981). They elucidate 30 separate standards and are presented in four groups that correspond to four main concerns about any evaluation: its utility, feasibility, propriety, and accuracy. They provide a framework for conducting good evaluation, espe-

TABLE 2

Analysis of the Relative Importance of 30 Standards in Performing Tasks in Evaluation

Table 4 ANALYSIS OF THE RELATIVE IMPORTANCE OF 30 STANDARDS IN PERFORMING 10 TASKS IN EVALUATION

	1 Deciding to do a Study	2 Clarifying Purpose	3 Political Viability	4 Contract	5 Staff Study	6 Manage Study	7 Collect Data	8 Analyze Data	9 Report Findings	10 Apply Results
A1 Audience Identification	X	X	X	X		X			X	X
A2 Evaluator Credibility	X		X	X	X	X	X			
A3 Information Scope				X			X		X	
A4 Valuational Interpretation		X	X				X	X	X	X
A5 Report Clarity						X			X	
A6 Report Dissemination			X	X		X			X	X
A7 Report Timeliness				X					X	
A8 Evaluation Impact	X	X	X						X	X
B1 Practical Procedures			X				X	X		
B2 Political Viability	X		X	X	X	X	X			X
B3 Cost Effectiveness	X	X				X				
C1 Formal Obligation	X		X	X		X	X			X
C2 Conflict of Interest	X	X	X	X	X	X				X
C3 Full and Frank Disclosure				X					X	
C4 Public's Right to Know			X	X					X	X
C5 Rights of Human Subjects			X	X		X	X			X
C6 Human Interaction			X			X	X			
C7 Balanced Reporting							X		X	X
C8 Fiscal Responsibility			X	X		X				
D1 Object Identification	X	X		X			X	X	X	
D2 Context Analysis	X	X					X	X	X	X
D3 Described Purposes	X	X	X	X		X	X		X	X
D4 Information Sources			X				X		X	
D5 Valid Measurement							X			
D6 Reliable Measurement							X			
D7 Systematic Data Control						X	X			
D8 Quantitative Analysis								X		
D9 Qualitative Analysis								X		
D10 Justified Conclusions		X						X	X	X
D11 Objective Reporting			X		X				X	

Source: Stufflebeam, D. L., & Madaus, G. (1983). The *Standards for Evaluation of Educational Programs, Projects, and Materials*: A descriptive summary. In Madaus, G. F., Scriven, M., & Stufflebeam, D. L. (Eds.). *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 395-404). Boston: Kluwer-Nijhoff. Used with permission.

cially at key checkpoints in any evaluation process. The *Standards for Evaluations of Educational Programs, Projects, and Materials* are presented in a way that facilitates their utility as guides to making decisions about evaluations addressed in this monograph. They can be especially helpful in deciding issues about using CRTs in program and curriculum evaluation.

As an additional aid to making useful evaluation decisions, 10 tasks in any evaluation are analyzed for their relative importance to the 30 standards. This information is presented in Table 2. Studying the matrix for information in Table 2 can be a good point at which to begin an exploration of evaluation issues.

Summation

A Blending of Information

A major goal for this monograph is to provide readers with information important to making good decisions about using CRTs in program and curriculum evaluation. The considerations in making such decisions are manifold: Decisions need to be made about programs and curricula, alternatives for contemporary approaches to evaluation must be discussed, and the instrumentation of any particular design should be reflected upon. It would be an oversimplification to think of the process for using a CRT in a program or curriculum evaluation as a puzzle with interlocking, discrete pieces. Rather, the interaction of the two principal components—the CRT and the evaluation model—is a blending process, akin to mixing different primary colors on a posterboard: Blue and yellow

FIGURE 7. *Instructional step number 5*

Determine whether the evaluation model you are considering allows for the instrumentation of a CRT by asking these questions:

- What is the primary purpose for this evaluation model?
- How was it constructed?
- What is the specificity of the information yielded?
- What is the generalizability of the information?
- What use will be made of the evaluation results?

are lost as distinct colors as a beautiful green emerges. So, too, with the information here. When properly considered by the criteria and methods suggested in this monograph, a CRT may provide substance to an evaluation framework, working together to produce information of value. Each circumstance will be new and unique, requiring persons of creativity, ingenuity, and good judgment to properly understand and appreciate the use of CRTs in program and curriculum evaluation. It is hoped that by following the instructional steps outlined in this monograph, one may be able to gain information of value by correctly using CRTs in program and curriculum evaluation.

Annotated References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
Specifies widely accepted industry standards for judging educational and psychological tests.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.
Widely used text in many introductory measurement classes.

Berk, R. (Ed.). (1980). *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press.
A collection of uneven essays about issues in criterion-referenced measurement.

Berk, R. (Ed.). (1984). *A guide to criterion-referenced test construction*. Baltimore, MD: Johns Hopkins University Press.
A collection of several important essays about some issues in criterion-referenced measurement.

Joint Committee. (1981). *Standards for evaluation of educational programs, projects, and materials*. New York: McGraw-Hill.
Describes standards for judging the merits of evaluation of educational programs and curriculum.

Madaus, G. F., Scriven, M., & Stufflebeam, D. L. (Eds.). (1983). *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 395–404). Boston: Kluwer-Nijhoff.
A collection of evaluation essays describing many commonly used models.

Nitko, A. J. (1980). A scheme for classifying and distinguishing criterion-referenced tests. *Review of Educational Research, 50*, 461–485.
An important article for classifying criterion-referenced tests.

Popham, W. J. (1984). Specifying the domain of content or behaviors. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 29–48). Baltimore: The Johns Hopkins University Press.
A good essay as introduction to the notion of clearly specifying a domain of content before attempting its assessment.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*(9), 679–682.
Strong argument on the merits of using assessment for defining paths to learning, rather than merely summative judgments.

Steele, S. (1973). *Contemporary approaches to program evaluation*. Washington, DC : Capitol Publications.
A practical collection of evaluation approaches particularly useful to practitioners.

Stufflebeam, D. L., & Madaus, G. (1983). Standards for evaluation of educational programs, projects, and materials: A descriptive summary. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 395–404). Boston: Kluwer-Nijhoff.
An important essay discussing frameworks for evaluation.

Student's Self-Test

This "Student's Self-Test" is designed to accompany the ITEMS module "Using CRTs in Program/Curriculum Evaluation." It presumes you have read the monograph and will help you determine whether you understood the information presented. As you consider the questions in this test you may need to refer to the monograph. Also, you may find more complete information presented on particular points in one of the references cited in the "Annotated References" section. Finally, your instructor may obtain an "Instructor's Guide" (see "Teaching Aids" ordering information). Although your own understanding of the information is not predicated on the information in the Instructor's Guide, it may be worthwhile to review the content of the monograph with someone more experienced than you are in using CRTs in program and curriculum evaluation.

1. Is this statement true or false: The relationship between evaluation and CRTs is akin to the conundrum "Which came first, the chicken or the egg?"

The question is *false* because it is clearly the framework for evaluation that prescribes or prohibits the use of a CRT in an evaluation.

2. Name the three types of program environments.

The three types of program environments are (a) regular district programs, (b) special compensatory programs, and (c) discretionary programs.

3. What are six criteria pertinent to installing CRTs in evaluation that are helpful in deciding which approach for evaluation may be appropriate?

Six such criteria are (a) evaluations that will help make major decisions and guide overall program management, (b) evaluations for examining the impact and larger results of programs, (c) evaluations that guide the organization and use of program components, (d) evaluations for viewing the program through the eyes of the participant, (e) evaluations for examining the results of instruction, and (f) evaluations that can serve specific purposes.

4. What are three requisite features for CRTs?

Three requisite features for CRTs are (a) clearly defined performance standards for measurement, (b) test items constructed specifically to address the intended performance standards, and (c) scores can be interpreted in terms of an individual's achievement of the specified performance standards.

5. Describe two general categories for CRTs.

CRTs may be classified as having well-defined and ordered domains or they may possess well-defined but unordered domains.

6. What is the publication that specifies widely accepted industry standards for judging educational and psychological tests?

The *Standards for Educational and Psychological Testing*.

7. What is the publication that describes standards for judging the merits of evaluation of educational programs and curriculum?

The *Standards for Evaluations of Educational Programs, Projects, and Materials*.

Instructor's Guide

to

USING CRTs IN PROGRAM/CURRICULUM EVALUATION

The material presented in this Instructor's Guide is designed to assist one in guiding learning experiences for students who study the monograph Using CRTs in Program/Curriculum Evaluation. The monograph describes how CRTs can aid the process of making evaluative judgments about educational programs and curricula. It is important to emphasize to learners from the outset that evaluation is a broadly based process encompassing a number of methodologies, each of which may employ a variety of tools, the CRT being but one of these aids. The enormous importance of CRTs in program and curriculum evaluation, however, warrants the attention of an entire monograph.

A principal goal of the monograph is to describe concepts and procedures in terms that are instructionally illuminating. The reader is guided to identify and examine particular points at which decisions must be made about how, when, and why CRTs may aid the evaluation process. These points are each identified as "An Instructional Step" and presented in separate boxes with pertinent guiding questions. All of the "An Instructional Step" boxes are included with this Instructor's Guide in an enlarged format to facilitate producing visual aids or overhead transparencies. Also, a number of checklists, tables, and charts are included with

this Instructor's Guide to serve as quick reference to particular instructional items. Some--but not all--of these checklists, tables, and charts are incorporated into the monograph. The additional ones may support the instruction of those presented to the student in the monograph itself. The following list presents the titles for each. Those that are also included in the monograph are noted with an asterisk.

*• Table 1 A SYSTEM FOR CLASSIFYING PROGRAM AND CURRICULUM EVALUATION APPROACHES

*• Table 2 A SCHEME FOR CLASSIFYING AND DISTINGUISHING CRITERION-REFERENCED TESTS

*• Table 3 A SUGGESTED OUTLINE FOR TEST EVALUATION

*• Table 4 ANALYSIS OF RELATIVE IMPORTANCE OF 30 STANDARDS IN PERFORMING 10 TASKS IN EVALUATION

*• AN INSTRUCTIONAL STEP #1

*• AN INSTRUCTIONAL STEP #2

*• AN INSTRUCTIONAL STEP #3

*• AN INSTRUCTIONAL STEP #4

*• AN INSTRUCTIONAL STEP #5

• Table 5 CATEGORIES OF CRITERION-REFERENCED TESTS BASED ON WELL-DEFINED AND ORDERED DOMAINS

• Table 6 CATEGORIES OF CRITERION-REFERENCED TESTS BASED ON WELL-DEFINED BUT UNORDERED DOMAINS

• Table 7 SUMMARY OF THE STANDARDS FOR EVALUATION OF EDUCATIONAL PROGRAMS, PROJECTS, AND MATERIALS.

Conciseness is a further aim of the monograph, one consequence of which is that several important concepts are only cursorily described or alluded to in passing. This fact should be pointed out to the student. When particular points are questioned, follow-up study may be required. One avenue to explore for follow-up study is the References section cited at the end of the monograph. A list of references important to a more complete understanding of how CRTs may be used in program and curriculum evaluation appears at the end of the volume. This list is expanded from the list of citations included with the Instructional Module and it does not included annotations (as does the Instructional Module).

There is a variety of audiences for whom this information may be useful, including at least six distinct groups:

- classroom teachers, counselors, parents, and students
- designers of curricula and planners of instruction
- school system administrators and boards of education
- educational research specialists and other social scientists
- legislators and other public officials
- news media and the general public

Each of these audiences needs different information for their particular decisions. For example, parents want information about a single child's performance--their own son or daughter--while news media are more interested in scores reflecting an entire group's performance. Sometimes the data

may be reported identically for various audiences but will be viewed from different perspectives. For example, students' performance on a particular set of objectives may be useful data to classroom teachers as well as to designers of curricula, but the classroom teacher will typically see the data in terms of individual students about whom judgments will be tempered with knowledge from other sources. A teacher will probably be familiar with a particular student's personality and have some notion of the background and environments which make up the student's world. While designers of curricula are sensitive to social contexts for their programs, they are not typically aware of factors such as motivation, previous experience with similar subject matter, or methods of instruction, that affect a particular's student's performance. Recognizing this diversity of needs for evaluation information is critically important to understanding and appreciating CRT use in program and curriculum evaluation. This fact should be emphasized when working with students.

Finally, it is useful to point out to students why this information is important. It is important because the information presented in this monograph will assist one in understanding and appreciating how CRTs may be used in program and curriculum evaluation. The distinction between <u>understanding</u> and <u>appreciating</u> is this regard is not merely a semantic subtlety. <u>Understanding</u> the use of CRTs is a mechanical step that entails sufficient familiarity with

evaluation methods sufficient to gauge how CRTs may be correctly installed in an evaluation effort as well as a knowledge of psychometric techniques necessary to construct a valid and reliable instrument. Appreciation in this case is a far more subtle matter. Appreciation implies a sympathetic dimension: a sensitivity to the people involved and an awareness of consequences. Appreciation has implications for a correct interpretation of results. If this monograph is successful, it describes the skills for understanding as well as the knowledge necessary for appreciating CRT use in program and curriculum evaluation. The instructor should emphasize to students that only when one is armed with both an understanding and an appreciation can the activities suggested here be meaningful.

Expanded (unannotated) References

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author.

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Macmillan.

Berk, R. (Ed.). (1984). A guide to criterion-referenced test construction. Baltimore: Johns Hopkins.

Berk, R. (Ed.). (1981) Educational evaluation methodology: The state of the art. Baltimore: Johns Hopkins.

Berk R. (Ed.). (1980). Criterion-referenced measurement: The state of the art. Baltimore: Johns Hopkins.

Cronbach, L. J. & Associates. (1980). Toward reform of program evaluation: Aims, methods, and institutional arrangements. San Francisco: Jossey-Bass.

Gronlund, N. A. (1985). Measurement and evaluation in teaching (5th ed.). London: Macmillan.

Joint Committee. (1981). Standards for evaluation of educational programs, projects, and materials. New York: McGraw-Hill.

Madaus, G. F., Scriven, M., & Stufflebeam, D. L. (Eds.). (1983). Evaluation models: Viewpoints on educational and human services evaluation. Boston: Kluwer-Nijhoff.

Sax, G. (1980). Principles of educational and psychological measurement and evaluation. Belmont, CA: Wadsworth.

Steele, S. M. (1973). Contemporary approaches to program evaluation: Implications for evaluating programs for disadvantaged adults. Washington, DC: Capitol Publications.

Worthen, B. R., & Sanders, J. R. (1973). Educational evaluation: Theory and Practice. Worthington, OH: Charles A. Jones.

Table 1 A SYSTEM FOR CLASSIFYING PROGRAM/CURRICULUM EVALUATION APPROACHES

Evaluation of results

A. Results concerned with attainment of objectives

1. Tylerian Models
2. Instructional Models (e.g., Goal Referenced, etc.)
3. Program Objectives Approaches (e.g., Suchman approaches)

B. Results concerned with evaluation of outcomes and effects

1. Goal Free Evaluation (e.g., Scriven)
2. Research Models
3. Accountability
4. Multiple Change Approach

C. Evaluation of programs as functioning, producing systems

1. Patterns of program systems

a. Evaluation as input into decision making *(e.g., CIPP Model)*

1. Major or general decisions

a. Decision Centered Evaluation
b. Differential Evaluation

2. Specific decision situations (selection of materials, setting priorities, etc.)

a. Developmental Models
b. Discrepancy Evaluation

b. Evaluation of program parts

1. Actual component approach
2. Action-Impact Approach
3. Management models (e.g., PERT)

2. Patterns of program systems as mirrored in evaluation

a. Evaluation – kinds of data, types of activities

1. Countenance of Evaluation
2. System Role Model
3. Means-End Hierarchy

b. Evaluation processes

1. General processes

a. Appraisal Model
b. Natural Process Models

2. Specifically applied processes

a. Monitoring Evaluation
b. Improvement Evaluation

Source: Steele, S. (1973). *Contemporary approaches to program evaluation.* Washington, DC: Capitol Publications. Used with permission.

Table 2 A SCHEME FOR CLASSIFYING AND DISTINGUISHING CRITERION-REFERENCED TESTS

	Well-defined and Ordered Domains	Well-defined But Unordered Domains	Ill-defined Domains	Undefined Domains
Basis for Test Development	Ordering based on judgements of the social or aesthetic quality of an examinee's product or performance	Specifying the stimulus properties of the items to be included in the domain	Poorly articulated behavioral objectives	No attempt to define a domain to which test performance is referenced
	Ordering based on which level of difficulty or complexity a topic or subject is learned	Specifying the stimuli and the responses in the domain	Defining the domain only in terms of the particular items on the test	Using a cut-off score, but not defining a performance domain
	Ordering based on degree of proficiency with which a complex skill is performed	Specifying the "diagnostic" categories of the domain		
	Ordering based on prerequisite sequences for acquiring an intellectual or psychomotor skill	Specifying the abstractions, traits, or constructs that define the domain		
	Basis for Test Development	Other ways of specifying the domain are possible		
	Ordering based on an empirically defined latest test			
	Ordering on other bases is possible			

Source: Nitko, A. J. (1980). A scheme for classifying and distinguishing criterion-referenced tests. *Review of Educational Research, 50*, 461-485. Used with permission.

Table 3 A SUGGESTED OUTLINE FOR TEST EVALUATION

A. *General Information*
   - Title of instrument (including edition and forms if applicable)
   - Author(s) and publisher, date of publication
   - Time required to administer
   - Cost

B. *Brief Description of Purpose and Nature of the Instruments*
   - General type, nature of content
   - Population for which designed
   - Subtests and separate scores, types of items

C. *Practical Evaluation*
   - Qualitative features (design, ease of use, attractiveness, etc.)
   - Ease of administration, clarity of directions
   - Scoring procedures
   - Administrator qualifications and training
   - Face validity and examinee rapport

D. *Technical Evaluation*
   1. *Norms*
      - Type of norm-based scores
      - Standardization sample (size, representativeness)
   2. *Reliability*
      - Type and procedure including size and nature of samples employed
      - Scorer reliability if applicable
      - Equivalence of forms
      - Long-term stability
   3. *Validity*
      - Specification of variable supposed to be measured
      - Appropriate types of validation procedures (content, criterion-related, construct)
      - Specific procedures followed in assessing validity and results obtained
      - Size and nature of samples employed

E. *Reviewers' Comments*
   - From *Mental Measurements Yearbooks*, journal reviews, or other sources

F. *Summary Evaluation*
   - Major strengths and weaknesses across all categories

Source: Anastasi, A. (1982). *Psychological Testing* (5th ed.). New York: MacMillan. Used with permission.

Table 4 ANALYSIS OF THE RELATIVE IMPORTANCE OF 30 STANDARDS IN PERFORMING 10 TASKS IN EVALUATION

	1 Deciding to do a Study	2 Clarifying Purpose	3 Political Viability	4 Contract	5 Staff Study	6 Manage Study	7 Collect Data	8 Analyze Data	9 Report Findings	10 Apply Results
A1 Audience Identification	X	X	X	X		X			X	X
A2 Evaluator Credibility	X		X	X	X	X	X			
A3 Information Scope				X			X		X	
A4 Valuational Interpretation		X	X				X	X	X	X
A5 Report Clarity						X			X	
A6 Report Dissemination			X	X		X			X	X
A7 Report Timeliness				X					X	
A8 Evaluation Impact	X	X	X						X	X
B1 Practical Procedures			X				X	X		
B2 Political Viability	X		X	X	X	X	X			X
B3 Cost Effectiveness	X	X				X				
C1 Formal Obligation	X		X	X		X	X			X
C2 Conflict of Interest	X	X	X	X	X	X				X
C3 Full and Frank Disclosure				X					X	
C4 Public's Right to Know			X	X					X	X
C5 Rights of Human Subjects			X	X		X	X			X
C6 Human Interaction			X			X	X			
C7 Balanced Reporting							X		X	X
C8 Fiscal Responsibility			X	X		X				
D1 Object Identification	X	X		X			X	X	X	
D2 Context Analysis	X	X					X	X	X	X
D3 Described Purposes	X	X	X	X		X	X		X	X
D4 Information Sources			X				X		X	
D5 Valid Measurement							X			
D6 Reliable Measurement							X			
D7 Systematic Data Control						X	X			
D8 Quantitative Analysis								X		
D9 Qualitative Analysis								X		
D10 Justified Conclusions		X						X	X	X
D11 Objective Reporting			X		X				X	

Source: Stufflebeam, D. L., & Madaus, G. (1983). The *Standards for Evaluation of Educational Programs, Projects, and Materials*: A descriptive summary. In Madaus, G. F., Scriven, M., & Stufflebeam, D. L. (Eds.). *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 395-404). Boston: Kluwer-Nijhoff. Used with permission.

Table 5 CATEGORIES OF CRITERION-REFERENCED TESTS BASED ON WELL-DEFINED AND ORDERED DOMAINS

Basis for Scaling or Ordering the Defined Domain of Behavoir [a]	Examples [b]
Judged social or esthetic quality of the performance	Rev. George Fisher's Scale Books (Chadwick, 1864) E. L. Thorndike's Handwriting (1910) and Drawing (1913) Scales
Complexity or difficulty level of the subject matter	Ayree's Spelling Scale (1915) Glaser's Criterion-referenced Measures I (1962, 1963)
Degree of proficiency with which complex skills are performed	Cox and Graham's Arithmetic Scale (1966) Harvard-Newton English Composition Scales (Ballou, 1914) Glaser's Criterion-referenced Measures II (1962, 1963) Perhaps certain sports events or physical fitness tests (1971)
Prerequisite sequence for acquiring intellectual and psychomotor skills	Gagné's Learning Hierarchies (1962) Piagetian Development Scales (Gray, 1978) Infant Development Scales (Uzgiris & Hunt, 1966)
Location on an empirically defined latent trait	Connolly, Nachtman, and Pritchett's arithmetic tests (1971) Other tests built with latent trait models (e.g., Rasch, 1960, or Birnbaum, 1968), provided they are referenced to well-defined and ordered domains of behavior.

Source: Nitko, A. J. (1980). A scheme for classifying and distinguishing criterion-referenced tests. *Review of Educational Research*, *50*, 461-485. Used with permission

a Other bases for scaling are possible.

b Examples are meant to be illustrative rather than representative or exhaustive.

Table 6 CATEGORIES OF CRITERION-REFERENCED TESTS BASED ON WELL-DEFINED BUT UNORDERED DOMAINS

Basis for Delineating the Behavior Domain[a]	During Test Development Emphasis is Placed on:	Examples[b]
Stimulus Properties of the Domain and the Sampling Plan of the Test	Defining content and content strata	Starch's English Vocabulary Test (1916) Ebel's Content-standard English Vocabulary Test (1962)
	Specifying stimulus properties of item domains	Hively's Item Forms (Hively, Patterson & Page, 1966)
		Osburn's Item Forms (1968)
Verbal Statements of Stimuli and Rèsponses in Domain	Behavioral objectives with or without the cut-off score ("criterion") specified	Tests based on Mager's Type of Objectives (1962) Curriculum Embedded Tests of IPI Mathematics
		Popham and Husek's Criterion-referenced Testing (1969)
		Harris and Stewart's Criterion-referenced Testing (1971)
	Elaborated description of behaviors and stimuli	Popham's Criterion-referenced Tests (1975, 1978)
		IOX Test Specifications (Popham, 1978, 1980, 1981)
"Diagnostic" Categories of Performance	Identifying entry-level behaviors	Hunt and Kirk's Tests of School Readiness (1974)
	Identifying behavior components missing from a complex performance	Tests built on Resnick's Components Analysis (Resnick, Wang, & Kaplan, 1973)
		Gagné's Two-Stage Testing (1970)
	Identifying and categorizing erreneous responses	Glaser, Darmin, and Gardner's "Tab-Item" Technique (1954)
		Hsu's Computer-Assisted Diagnostic Tests (Hsu & Carlson, 1972)
	Identifying erroneous processes	Beck's Blending Algorithm (Beck & Mitroff, 1972)
		Interviews to determine what processes were used in responding to test tasks
Abstractions, Traits, or Constructs	Specifying specific behaviors or categories of behaviors that delimit the abstraction,trait, or construct	Tests based on the *Taxonomy of Educational Objectives* (Bloom, 1956)

Adapted from: Nitko, A. J. (1980). A scheme for classifying and distinguishing criterion-referenced tests. *Review of Educational Research*, *50* , 461-485. Used with permission.

a Other bases for delineating are possible.

b Examples are meant to be illustrative rather than representative or exhaustive.

Table 7 SUMMARY OF THE STANDARDS FOR EVALUATION OF EDUCATIONAL PROGRAMS, PROJECTS, AND MATERIALS

A **Utility Standards**
The utility standards are intended to ensure that an evaluation will serve the practical information needs of given audiences. These standards are:

A1 **Audience Identification**
Audiences involved in or affected by the evaluation should be identified, so that their needs can be addresses.

A2 **Evaluator Credibility**
The persons conducting the evaluation should be both trustworthy and competent to perform the evaluation, so that their findings achieve maximum credibility and acceptance.

A3 **Information Scope and Selection**
Information collected should be of such scope and selected in such ways as to address pertinent questions about the object of the evaluation and be responsive to the needs and interests of specified audiences.

A4 **Valuational Interpretation**
The perspectives, procedures, and rationale used to interpret the findings should be carefully described, so that the bases for value judgments are clear.

A5 **Report Clarity**
The evaluation report should describe the object being evaluated and its context, and the purposes, procedures, and findings of the evaluation, so that the audiences will readily understand what was done, why it was done, what information was obtained, what conclusions were drawn, and what recommendations were made.

A6 **Report Dissemination**
Evaluation findings should be disseminated to clients and other right-to-know audiences, so that they can asses and use the findings.

A7 **Report Timeliness**
Release of reports should be timely, so that audiences can best use the reported information.

A8 **Evaluation Impact**
Evaluations should be planned and conducted in ways that encourage follow-through by members of the audiences.

B **Feasibility Standards**
The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal; they are:

B1 **Practical Procedures**
The evaluation procedures should be practical, so that disruption is kept to a minimum and that needed information can be obtained.

B2 **Political Viability**
The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.

Table 7 SUMMARY OF THE STANDARDS FOR EVALUATION OF EDUCATIONAL PROGRAMS, PROJECTS, AND MATERIALS (Cont'd.)

B3 **Cost Effectiveness**
The evaluation should produce information of sufficient value to justify the resources extended.

C **Propriety Standards**
The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results. These standards are:

C1 **Formal Obligation**
Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligation to adhere to all conditions of the agreement or formally to renegotiate it.

C2 **Conflict of Interest**
Conflict of interest, frequently unavoidable, should be dealt with openly and honestly, so that it does not compromise the evaluation processes and results.

C3 **Full and Frank Disclosure**
Oral and written evaluation reports should be open, direct, and honest in their disclosure of pertinent findings, including the limitations of the evaluation.

C4 **Public's Right to Know**
The formal parties to an evaluation should respect and assure the public's right to know, within the limits of other related principles and statutes, such as those dealing with public safety and the right of privacy.

C5 **Rights of Human Subjects**
Evaluations should be designed and conducted so that the rights and welfare of the human subjects are respected and protected.

C6 **Human Interactions**
Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation.

C7 **Balanced Reporting**
The evaluation should be complete and fair in its presentation of strengths and weaknesses of the object under investigation, so that strengths can be built upon and problem areas addressed.

C8 **Fiscal Responsibility**
The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible.

D **Accuracy Standards**
The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features of the object being studied that determine its worth or merit. These standards are:

D1 **Object Identification**
The object of the evaluation (program, project, material) should be sufficiently examined, so that the form(s) of the object being considered in the evaluation can be clearly identified.

D2 **Context Analysis**
The context in which the program, project, or material exists should be examined in enough detail so that its likely influences on the object can be identified.

Table 7 SUMMARY OF THE STANDARDS FOR EVALUATION OF EDUCATIONAL PROGRAMS, PROJECTS, AND MATERIALS (Cont'd.)

D3 **Described Purposes and Procedures**
The purposes and procedures of the evaluation should be monitored and described in enough detail so that they can be identified and assessed.

D4 **Defensible Information Sources**
The sources of information should be described in enough detail so that the adequacy of the information can be assessed.

D5 **Valid Measurement**
The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is valid for the given use.

D6 **Reliable Measurement**
The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the information obtained is sufficiently reliable for the intended use.

D7 **Systematic Data Control**
The data collected, processed, and reported in an evaluation should be reviewed and corrected, so that the results of the evaluation will not be flawed.

D8 **Analysis of Quantitative Information**
Quantitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.

D9 **Analysis of Qualitative Information**
Qualitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.

D10 **Justified Conclusions**
The conclusions reached in an evaluation should be explicitly justified, so that the audiences can assess them.

D11 **Objective Reporting**
The evaluation procedures should provide safeguards to protect the evaluation findings and reports against distortion by the personal feelings and biases of any party to the evaluation.

Source: Stufflebeam, D. L., & Madaus, G. (1983). The *Standards for Evaluation of Educational Programs, Projects, and Materials*: A descriptive summary. In Madaus, G.F., Scriven, M., & Stufflebeam, D. L. (Eds.). *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 395-404). Boston: Kluwer-Nijhoff. Used with permission.

An Instructional Step # 1

Determine in which program environment you may install a CRT for program or curriculum evaluation by asking these questions:

- Is this program a regular district program?
- Is this program a special compensatory program?
- Is this program a discretionary program?

An Instructional Step # 2

Determine from the above classifying criteria and from Table 1 a suggested approach to evaluation by following these steps:

- Review the criteria for classifying evaluation approaches as a prompt to asking yourself the relevant questions for your particular evaluation.

- Find the portion of the taxonomy that most nearly addresses your evaluation concerns.

- Read in the taxonomy a suggested approach to evaluation. [NOTE: if you are unfamiliar with the evaluation tool suggested, you may find a description of the model in one or more of the references cited in the References section of this monograph.

An Instructional Step # 3

Determine whether the measure you are considering is criterion-referenced by asking these questions:

- What is the primary purpose for which this test was constructed?
- How was it constructed?
- What is the specificity of the information yielded about the domain of instructionally relevant tasks?
- What is the generalizability of test performance information to the domain? [NOTE: this question is applicable to most, but not all, CRTs.]
- What use will be made of the obtained test information?

An Instructional Step # 4

Determine classificatory features of the CRT you are considering by asking these questions:

- Are the domains to be measured well-defined, ill-defined, or undefined? [Note: use the "Basis for Test Development" criteria listed in Table 2 to guide you.] If the domains are ill-defined or undefined, the measure is not a true CRT and is inappropriate for the present purposes.

- Is the domain ordered? If so, then find the basis for scaling or ordering the defined domain.

- Is the domain unordered? If so, then find the basis for delineating the behavior domain and the area for emphasis during test development.

An Instructional Step # 5

Determine whether the evaluation model you are considering allows for the instrumentation of a CRT by asking these questions:

- What is the primary purpose for this evaluation model?
- How was it constructed?
- What is the specificity of the information yielded?
- What is the generalizability of the information?
- What use will be made of the evaluation results?