STANDARDS FOR EVALUATING CRITERION-REFERENCED TESTS

Clinton B. Walker

CSE Report No. 103

January 1978

Center for the Study of Evaluation
University of California, Los Angeles
Los Angeles, California 90024

This Project was supported in whole or in part by the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Standards for Evaluating Criterion-Referenced Tests

Clinton B. Walker[1]

INTRODUCTION

This report presents a set of standards for judging the merits of criterion-referenced tests (CRTs). The standards were developed in connection with CSE Criterion-Referenced Test Evaluations (forthcoming). This resource book contains evaluative and descriptive reviews of sixty-four criterion-referenced tests, primarily in the basic skills areas.

In planning a system for evaluating CRTs, CSE decided to create one set of standards to apply to all tests. The need for test users to be able to compare different tests dictated the decision to have a uniform evaluative system. An initial pool of 70 possible standards was gathered by reviewing the professional literature and publishers' promotional materials. This large number was reduced by combining some possible judgments, eliminating factors that were relevant only to norm-referenced tests, and by treating some other judgments as descriptive rather than evaluative information. After extensive internal and external review, a set of 21

[1] The aid of project research staff is gratefully acknowledged. Russell Hunter gathered the large pool of standards, criteria, and dimensions which launched this effort. A formative tryout of the standards on a sample of tests was conducted by Russell Hunter, Karen Smith, Shonagh Neafsey, and Margeret Dotseth. In addition, critical external reviews of a draft of these standards were made by Jack C. Merwin (University of Minnesota), Albert H. Rouse, Jr. (Department of Research and Development, Cincinnati City School District), and Jason Millman (Cornell University). We took the liberty of adopting, interpreting, and in some cases resisting their good advice.

standards was produced. (See Table 1.) These fall into three groups: Measurement Properties, Appropriateness for Examinees, and Practicality.

There is one standard of overriding importance to a test buyer which could not be included: the degree of correspondence between the objectives of a test package and the objectives of any specific curriculum to be tested. In selecting a CRT system a test buyer will want to pick the one which is most relevant to local instruction. By selecting a test in this manner, one maximizes the likelihood that its scores will reflect what is actually being taught and actually being learned.

To some extent the use or function of a CRT will dictate the standards which are relevant for weighing its merits. The classroom management uses such as diagnosis, prescription, and progress monitoring call for a CRT to have some traits which are not important in using CRTs for program evaluation. For example, it is important to have reliability in measuring individual students when using CRTs for decisions about individuals, but reliability is needed only for larger units when evaluating programs. CSE emphasized the classroom management uses of CRTs in the standards dealing with (8) consistency, (14) curriculum cross-referencing (19) record keeping, and (20) decision rules.

CRTs can also be used effectively for program evaluation. One economical method for doing so involves sampling of objectives and items, so as not to have a lengthy test, and sampling of pupils as well. In this context the CRTs are not intended to support decisions about individuals, but are meant to give a survey of achievement on classroom level teaching objectives. Standards 8, 14, 19, and 20 mentioned above are not important in that context.

Table 1

21 Standards for Evaluating CRTs

IB. MEASUREMENT PROPERTIES: CONCEPTUAL VALIDITY

1. Description: How good (i.e., thorough and understandable) are the descriptions of the domains to be tested?
2. Agreement: How well do the test items match their objectives?
3. Representativeness: How adequately do the items sample their objectives?

IB. MEASUREMENT PROPERTIES: FIELD TEST VALIDITY

4. Sensitivity: Does conventional instruction lead to test-score gains?
5. Item Uniformity: How similar are the scores on the different items for an objective?
6. Divergent Validity: Are the scores for each objective relatively uninfluenced by other skills?
7. Lack of Bias: Are test scores unfairly affected by social group factors?
8. Consistency of Scores: Are scores on individual objectives consistent over time or over parallel test forms?

II. APPROPRIATENESS FOR EXAMINEES

9. Clarity of Instructions: How clear and complete are the instructions to students?
10. Item Review: Does the publisher report that items were reviewed or field tested for quality?
11. Visible Characteristics: Is the layout and print easily readable?
12. Ease of Responding: Is recording answers a problem for pupils?

III. PRACTICALITY

13. Informativeness: Does the test buyer have adequate information about the test before buying it?
14. Curriculum Cross-Referencing: Are the test objectives indexed to at least two series of relevant teaching materials?
15. Flexibility: Are many of the objectives tested at more than one level, and are single objectives easy to test separately?
16. Alternate Forms: Are parallel forms available for each test?
17. Test Administration: Are the directions to the tester clear, complete, and easy to use?
18. Scoring: Are both machine scoring and easy hand scoring available?
19. Record Keeping: Does the publisher provide record forms keyed to test objectives that are easy to use?
20. Decision Rules: Are well justified, easy to use rules given for making instructional decisions on the basis of test results?
21. Comparative Data: Are scores of a representative reference group of students given for comparing with your students' scores?

In selecting among tests the reader will thus need to adapt the CSE standards (or any others) to the specific use for which tests are being considered.

Ratings on the CSE standards are given in letter grades. Letter grades were used instead of numbers to encourage readers to weigh the standards according to their own needs, rather than to add up the ratings mechanically.

The standards that deal with the field-test validity of CRTs (standards 4 through 8) call for the reader to judge not the tests themselves, but instead publishers' reports about their tests. Since field-test data are not reported for many CRT packages as yet, the resulting ratings will be low. Care should be taken to distinguish between a low rating that comes from a lack of evidence from one that comes from a lack of quality. CSE maintains that using a relatively untested CRT system that is clearly related to the local curriculum is preferable to using a well field-tested norm-referenced test (NRT) that relates to the local curriculum in unrepresentative or unknown ways. In support of this position, Cronbach (1970, p. 152) has said that precision in test scores is useless if the skills measured by the test are not relevant to the intended decision. In the near future we can hope that the developers of NRTs will do a better job of keying their test items to clearly described, teachable objectives and that the developers of CRTs will carry out and report the needed validation studies.

Readers should view these standards as formative or developmental, because theory and practice in CRM are still emerging. For example, there is not general agreement yet on basic issues like the relevance of traditional validation procedures and how to set minimal performance levels. Although the standards presented below are debatable, they are not arbitrary; each one is introduced by a discussion of its importance in a section labelled Background.

I. MEASUREMENT PROPERTIES

A. Conceptual Validity

1. Description of the criterion

Background

Two major differences between CRTs and NRTs are that CRTs give a clearer description of the behavior to be tested and the validity of CRTs depends more heavily on the match between that description and the test items. Before discussing those points it is useful to note that a test score is not an end in itself: it is a sign or indicator of something more important. A score may give a prediction about the pupil's future performance, or it may give an estimate of how the test taker is likely to perform on the larger set of possible test items from which the items actually tested are drawn. In the latter case that pool of possible test items is called the criterion. A pupil's score on a CRT thus gives an estimate of how the pupil is likely to perform on the criterion, i.e., on the population of all such items.

One essential step in making the scores of a CRT meaningful is to describe the criterion pool of items clearly. First, a clear description enables teachers to teach to the skill or attitude that is described. By providing a practical target for instruction, the description makes the score on such a test useful for diagnosis and prescription. Secondly, a clear test description can help to demystify testing by telling consumers of test results just what was tested. The description thus gives meaning to the score. In most of the tests covered in the CSE Criterion-Referenced Test Evaluations the descriptions of the criterion behaviors take the form of instructional objectives. When "a test" is referred to, it means a group of items that provides a separate score. One CRT test form may thus contain several tests.

Since the items of a CRT are supposed to test the skill or attitude as set forth in the description of the criterion, the validity of a CRT depends on the extent to which the items actually fit the test description. This type of validity is often referred to as content validity, but that term is too narrow. Popham (1978) has suggested the phrase descriptive validity so as to apply not only to CRTs in the cognitive domain but also to those in the psychomotor and affective domains, where process or action may be more relevant than content. In order for descriptive validity to be judged or measured, it is necessary to have first a test description; without such a description it is not possible to tell what the criterion pool of test items is supposed to include and exclude (Linn, 1977). A clear description of the criterion is thus an essential link in determining that a CRT has descriptive validity.

What will an adequate CRT description contain? It will amount to a set of instructions to the test writer that prescribes the content, format, and mode of responding for all of the possible test items. Directions for making up multiple choice options, for scoring free responses, and for sampling items from the criterion item pool are also given. Much of this information goes beyond subject matter content, so the phrase content validity is again seen as too narrow to refer to the match or mismatch between CRT items and their test description.

It is obvious that such test descriptions are technical documents, too lengthy and detailed in their entirety to be useful either for planning instruction or for reporting grades. But they should include brief statements for teachers and parents in a form like behavioral objectives. The three approaches to careful test construction that have been proposed up to now are amplified objectives, domain specifications, and item forms. None of these is widely used as yet.

The standard

Level C - The test is described in terms that give little indication of the content, format, and response mode of the test items. Many different types of test items will fit a description this loose. Descriptions that do not mention the observable behavior of the test taker are at this level of completeness. These are scarcely criterion-referenced tests.

- Example: Odd and even numbers

  Problems: Format and response modes are not mentioned, nor is a plan for sampling from the possible items. Limits on the set of stimuli are not given.

- Example: Students will demonstrate their comprehension of passages of text by identifying their main idea.

  Problems: A verb phrase that describes the observable behavior of the test taker is needed. The type of text is not limited at all.

Level B - Content, format, and response mode are described, as in a behavioral objective. Rules for sampling items are not given, or there is so much slack in the limits of content, format, or response mode that somewhat different tasks could still fit the description. Tests based on such descriptions are objectives-based or objectives-referenced.

- Example: Given a paragraph of grade level text, students will demonstrate their comprehension by summarizing the main idea in writing in six words or less.

Problems: A sampling plan is needed to insure representativeness and avoid duplication. Maximum and minimum paragraph lengths need to be specified. The organization of ideas in the paragraphs needs to be stated so as to insure that main ideas are both present and of the desired types. Rules for scoring students' responses are needed.

Level A - Content, format, response mode, and sampling rules are described thoroughly enough so that a) different test writers would produce equivalent tests by following the description or b) for any test item or set of items it is clear whether they fall inside or outside the intended domain. It is important to make clear the boundaries of a set of test items as well as its core. The example given above at Level B does not serve this delimiting function. Item forms, amplified objectives, and domain specifications approach this clarity and completeness of description.

- Example: Figure 1 presents an example of an <u>item form</u>.

2. Agreement of items with their test descriptions

Background

The descriptions (Standard 1 above) are a test maker's intentions for constructing tests. It is still necessary to show that the intentions were carried out. Standards (2) and (3) deal with this issue. Standard (2) asks whether the test items are accurately described by the test description. If they are not, then the items test something else and the test is invalid. Technical terms that are used to refer to the concept of agreement include item-objective congruence, content validity, and descriptive validity.

The standard

Level C - No evidence of agreement is offered; or evidence is mentioned but not described in enough detail to evaluate; or evidence is described in detail but is flawed.

- Example: "Two subject matter specialists with experience in test construction were hired as consultants to judge the congruence of the items to their objectives. Both examined each objective and the set of items that was written for it. For each item the specialists marked whether it was congruent with the objective. Overall, 89% of the items were judged congruent by both judges. The other 11% of items were discarded from the final version of the test." (Imaginary description.)

Problems: By presenting the objectives only with their own test items the fictional test developer may have induced a bias in the specialists to expect agreement (congruence) and overlook disagreement. Also by examining a set of items together, judges may have been influenced by inter-item similarities.

Level A - Sound evidence of agreement is offered and described in enough detail to evaluate. The test developer gives a detailed account of either how the items were generated from the description of the criterion behaviors or how qualified judges confirmed the fit of the individual items to the description.

- Example: The previous example with the two problems corrected.

3. Representativeness of the items

Background

We are rarely interested in a test score for its own sake. Test scores are used as observable indicators of more important things that are difficult or impossible to observe. For example, students' scores on any achievement test are used to indicate their mastery of the total set of possible questions on the subject matter. It is rarely practical to test the total set. Likewise, a person's performance on a test of intelligence or personality is used as an indicator of how the person will act in more natural situations. In order for a test score to be an accurate indicator, the test items must be representatively selected. In other words, the items must be chosen in a way that lets us generalize from the test score to the intended total set of behaviors. If the selection process is biased, unplanned, or unrepresentative, then we cannot be sure what total set of behaviors the test score represents. Standard 3 deals with representativeness.

The standard

Level C - No account is given of how the test questions were chosen from the set of questions possible under this objective. Or, items were selected in a biased or unrepresentative fashion. Items are not representative if the item selection process systematically excluded those ones that failed to discriminate high and low scoring individuals in a group of students who have a common instructional background.

Level A - The test developer reports that the items were selected either randomly or, if there are components in the domain, by stratified random sampling.

B. Field Test Validity

Authorities in the field of CRM agree that the standards for conceptual validity are necessary for a good CRT. They do not agree, however, on the necessity for empirical (data-based) validation of CRTs. We take the position that the two types of validity are interdependent: both are necessary for confirming that a test measures what it claims to. Without validation by field trials, a test that appears to be conceptually sound may give measures that are not dependable (8), that do not reflect the relevant learning (4), that are of an unintended mixture of behaviors (5), that are affected by skills or attitudes other than the intended one (6), and that are biased (7). Without meeting the standards for Conceptual Validity, on the other hand, a test may be an unrepresentative measure (3) of the wrong criterion (2) or of no identified criterion at all (1).

4. Sensitivity to learning

Background

Students' scores on any test may or may not show the effects of real learning that has occurred. To the extent that they do, the test is said to be sensitive to learning. We are saying here that a test is better if it is sensitive. This standard for judging the merits of tests is not universally accepted, in part because it is usually called sensitivity to instruction. The objection is this: a test may not show any effects of instruction because the given instruction did not have any effect. Thus when a small sample of students in a field test does not appear on a post-test to have benefited from instruction, that result is not necessarily the fault of the test.

That objection is well taken as far as it goes. As consumers of tests, though, we need to know that the test does reflect the positive effects of instruction in a fair proportion of classrooms. If it does not, either the test is insensitive or the test content is not teachable. In either case such a test will not be useful.

A showing of sensitivity to learning under one form of instruction will not guarantee sensitivity to all forms of instruction. The test developer should describe the type(s) of teaching used in the field tests so that test buyers can decide if the test is sensitive to their type of teaching.

There are serious technical problems in measuring change, and there is not yet a consensus on how to prove a test's sensitivity. At this point in the history of CRM we have simply asked whether the test developer offers any evidence of a test's sensitivity that is free from the well established problems in measurement.

Level C - No information is given on the sensitivity of the test to student gains; or evidence is offered which suffers from well established problems in measurement, like those discussed in Campbell and Stanley (1963); or the gains are not statistically dependable; or the successful teaching method was not described.

Level A - The test has been found to reflect learning in a representative sample of students following an ordinary (in terms of time, intensity, and resources usually available for the particular subject) course of instruction. The course of instruction is clearly aimed at the criterion behaviors. The well established problems in measurement are not present in the study.

5. Item uniformity

Background

This standard deals with the evidence that a test measures a uniform, coherent skill or attitude. If the test does not, then it measures a mixture of things. A CRT that is a uniform measure is a better test, with the following exception: in some cases the definition of the criterion behaviors identifies different components or levels of difficulty. For example, a phonics test might deal with consonants, the different categories of consonants (e.g., stops, liquids, nasals, fricatives) being identified as components of that phonic skill. Such a test should show uniformity within each category, but not necessarily within the whole test of several categories. When such a test measures a mixture of things, the mix is planned. An accidental lack of uniformity results when the items unintentionally call for different skills or attitudes. It is a sign that the description of the criterion is defective, for the test does not measure what it purports to measure.

Item uniformity is important for a second reason. We want each test score to correspond to a specific amount of learning or skill. If the correlation among items for an objective is low, then any given score on the objective (like 80% correct) will reflect different levels of achievement depending on which specific items were answered correctly. In other words, as item uniformity increases, the scores for that objective become more exact indicators of pupils' level of achievement.

Uniformity or coherence of a CRT is shown by measures of the extent to which all the items function alike. The more that students' scores on one test item are similar to their own scores on the other items, the more uniformity the test has. In classical statistical theory factor analysis, inter-item correlations, and part-whole correlations give measures of uniformity.

The standard

Level C - No numerical evidence of item uniformity is given; or only judgmental evidence is given; or the data are for several objectives taken together.

Level A - At this early stage in the history of CRM, any numerical evidence of item uniformity will be accepted. The data must be based on students' responses to the test items, and must be based on data for individual objectives.

6. Divergent validity

Background

This standard deals with whether the scores on a test are relatively uninfluenced by achievements or attitudes that the test is not supposed to be measuring. If the scores on the test are uninfluenced by other unintended factors, then it is a test of something distinct and has divergent validity. Consider this example: students' scores on a test of reading comprehension may correlate highly with their scores on a vocabulary test. It is questionable then that the comprehension test is measuring anything separate from vocabulary, and it lacks divergent validity. In order for a math test to have divergent validity, the language in it must be simple enough so that pupils' errors are not reading errors.

Divergent validity, or separateness, can be confirmed by traditional methods like factor analysis, low correlations between measures of supposedly separate behaviors, or by experimental research showing that one test responds to a treatment while others do not.

The standard

Level C - No evidence of divergence is offered, or the evidence is not detailed enough to judge. Evidence of contamination is offered (e.g., high correlations of CRT scores with either I. Q. scores or scores of verbal aptitude).

Level A - Evidence of divergent validity is given that shows the CRT's scores to be independent of scores on tests of other supposedly unrelated achievements or attitudes.

7. Lack of bias

Background

A test is biased for a given group of students if it does not permit them to show their true achievement or attitude as completely as it permits other groups to do so. A test that does not do this is invalid for that group The subject of bias is surrounded with controversy, in part because social justice for large numbers of students is at stake.

This standard is concerned with how different groups of students, for example different ethnic groups, perform on a test. It is not concerned with the content of test questions. Bias has been common enough in testing that we cannot assume it to be absent from current tests. Hence we are asking for a showing of lack of bias to confirm a test's validity for major social groups.

The standard

Level C - No evidence of lack of bias is offered, or evidence is offered but not persuasive. A difference in the average scores of ethnic or other groups by itself will not be considered evidence of bias.

Level A - Evidence of lack of bias is offered for at least two of the following groups: women, blacks, and students from Spanish speaking backgrounds. Lack of sizable item by group interactions is one form of evidence. A second is the similarity across groups of the other data for empirical validity, standards (4)-(8).

8. Consistency of scores

Background

A test is better if the difference in a student's scores on two occasions are due to a real change in achievement, or, for affective measures, in attitude. If a student's scores change due to the vagueness of the instructions, variations in testing conditions, or other factors aside from real learning, then the test's scores are not consistent. Changes in scores due to irrelevant factors make the scores on any one occasion suspect. The more that a test's scores reflect real learning, and not these irrelevant factors, the more consistent we say it is.

Consistency measures used with norm-referenced tests include test-retest reliability and alternate form reliability. It is not clear whether these same statistics will be suitable for CRTs, so we are using the broader term consistency.

Consistency data are necessary to show that a test's scores are dependable, but not many such studies have been done yet on CRTs. In principle consistency may vary over a wide range, but current CRTs differ more on whether they report stability data at all than on the values reported. Thus at this point in the history of CRM we have asked only whether any such data are reported. It is a positive step in test development to measure consistency and a step toward truth in packaging to report the results.

The standard

Level C - No consistency data are given.

Level A - Data are reported on the consistency of students' scores. Either consistency of individuals' scores over repeated testing or consistency of individuals' scores on different forms of the test will be credited.

II. APPROPRIATENESS FOR THE INTENDED EXAMINEES

The effects of the following four factors would show up in the validity and consistency data for a test. Because little information is available yet on the measurement properties of CRTs, and because the following four factors may cause problems in giving a test, they are treated separately here.

9. Clarity of instruction to students

Background

The instructions to students must describe all aspects of the task in language that is suited to the intended age or grade levels. A sample item that is typical and clear should be given both for practice and clarification.

Level C - The language of the instructions is too advanced or too basic; or instructions are incomplete or hard to follow; or a sample item is not given.

Level B - Either the instructions or the sample item are lacking.

Level A - The instructions are clear and complete to the intended test takers and a sample item is provided.

10. Item Review

Background

Test items are appropriate if they are understandable, have at least one correct answer, give credit for all correct answers, do not give away the correct answer, and are otherwise free from technical flaws. One kind of evidence is considered here, namely test developers' reports of item quality control review.

The standard

Level C - The test developer offers no evidence that item quality was checked apart from the process of original item generation.

Level A - The test developer reports that item quality was reviewed independently of item writing.

11. Visible characteristics of test materials

Background

The visible characteristics of test materials should make it easy for students at the intended levels to use the information therein. Tests of all objectives were examined for the details of layout, organization, and clarity mentioned in the standard below.

The standard

Level C - $\geq$ 10% of the objectives have one or more of these flaws: print or pictures unclear, items too close together, stems and responses not clearly grouped, sequence of items easy to lose, little blank space for math work, page cluttered, item numbers not easy to pick out, information needed to answer one question unnecessarily spread out.

Level A - < 10% of the objectives have such flaws.

12. Ease of responding

Background

A test should be set up so that students' scores are not affected by difficulties in recording their answers. Answer sheets or other spaces for responding are judged for not only their physical attributes, as in standard (11), but also whether answer spaces are large enough.

The standard

Level C - Answer sheets or other response spaces for $\geq$ 10% of the objectives have one or more of these flaws: unclear print, items too close together, item numbers not easy to pick out, answer spaces too small.

Level A - < 10% of objectives have such flaws in the response materials.

III. PRACTICALITY

13. Informativeness of materials for the prospective buyer

Background

Some testing systems make it easier for the prospective buyer to make an informed choice to buy or not to by providing complete, easy to use information on their system. Two stages in test purchase are usual: deciding to order sample materials and deciding to order the testing package itself. Since the presence and quality of technical information on test development is covered above in standards (2) - (8), it will not be counted again here. The issue here is whether the prospective buyer knows what the testing package will consist of before investing in it. This standard is more important in weighing the more costly CRT systems, where the prospective user will be less willing to buy the system before deciding whether to use it.

The standard

Level C - Promotional materials like flyers and catalogs, do not give enough information for deciding whether to order specimen sets. Or, specimen sets or sample pages and instructions are not offered; tests can be purchased only in multiple copies. Or, a complete listing of the test's objectives is not provided before purchase. Or, information of ordering of original and replacement materials is not clear and complete. Or, replacement materials may not be ordered separately. Or, information on returning unused materials is vague. Or, the following types of information are not available without buying the testing package: what the instructions to students and test user are, what the physical characteristics of the test are, how and where students record their answers, how many separate test forms there are, what sorts of decision standards and comparative data are provided, how long the tests are supposed to take, and whether special training is needed to give or interpret the test.

Level A - All of the information above is available before the prospective user buys the testing system. Or, the whole system is available on approval.

14. Curriculum cross-referencing

Background

A testing package is easier to coordinate with local curriculum and instruction if it includes an index relating its specific tests to specific teaching materials. Such an index can be used to guide test selection and to help teachers locate alternative instructional materials. For either purpose the user will have to verify that the indexed instructional materials adequately cover the same skills as the respective tests and the local curriculum.

The standard

Level C - No curriculum cross-referencing is provided.

Level A - Indexing of the tests to two or more publishers' instructional materials is provided in detail (e.g., specific units in specific texts).

15. Flexibility of choosing objectives

Background

A test or testing system is adaptable to a range of local needs, like individualization, if it covers a variety of objectives and tests them on separate forms. A testing system which tests mixtures of objectives together does not give the local user as much control over testing. Also, a system which gives tests of the same core objectives for more than one grade level is more flexible. Such a system does not have the same test items on forms that differ only in the level marking, but has tests of the same skills with content and illustrations suited to the different levels.

Note that it is not fair to compare large scale testing systems with smaller ones that do not try to cover the same range of skills or grades. If the user is looking for a more specific test, this criterion may not be relevant. Also the cost of the flexibility will be an important consideration to the test buyer.

The standard

Level C - The test or system provides a narrow range of objectives and prints several of them together on the same test form. Core objectives are available in materials appropriate to only one grade level.

Level B - One of the above features is missing: variety, separate forms, or grade level flexibility of core objectives.

Level A - All of the features are present.

16. Alternate forms

Background

When a testing system has alternate forms the user can give independent retests to the same students. If retesting is done with the same form that was used for the original test, students' scores are likely to be influenced not only by their learning of the subject matter but also by sepcific memory of the first testing. This latter influence invalidates the retest scores. With alternate forms, pre- and posttesting or repeated posttesting can be carried out without this invalidating carry-over effect.

The standard

Level C - Only one form is available for each test.

Level A - Two or more forms with non-overlapping sets of items are available for each test.

17. Test Administration

Background

A test is more practical if the instructions to the examiner are clear, complete, and well organized. With good instructions the testing is not only easier but also the testing conditions are more uniform.

The standard

Level C - Instructions to the examiner are hard to find or follow. They are vague, ambiguous, not complete, not all in one place, or not logically ordered. Or, the copy in the manual is unclear.

Level A - Insttuctions leave little room for misunderstanding by the examiner and are complete and easy to use.

18. Scoring

Background

A test is more practical if it can be scored easily and objectively and if the test user is not limited to one method of scoring. Hand scoring is easy if scoring templates or other well organized keys are provided.

The standard

Level C - Hand scoring is difficult, or arbitrary, or requires special training. Or, scoring requires the expense of special machines on site or the delay of sending students' responses out for scoring.

Level B - Hand scoring only, but it is objective and easy.

Level A - Both machine and easy, objective hand scoring options are available.

19. Record Keeping

Background

Good records of student performance are an important part of classroom management and of meeting accountability requirements. CRT systems have the potential for making record keeping a big job because they often have large numbers of objectives. A testing system is more practical when it has forms for recording students' test scores that are easily keyed to the objectives, easy to maintain, and easy to interpret.

The standard

Level C - Either teachers must create their own record forms or the testing system's forms are not easily keyed to the objectives, easy to maintain, or easy to interpret.

Level A - Usable forms for record keeping are provided.

20. Decision rules

Background

Tests are devices for making decisions about students. Tests should be constructed in a way that allows decisions to be made with confidence and ease. The information for decision making should be easy to find, easy to use, and well justified. Although the choice of cutting scores for passing or mastery should be up to the local test user, the publisher should give an indication of the consequences of choosing different cutoffs.

Relative costs and gains will affect the choice of a cutting score. Where a prerequisite skill is being tested it may be preferable to hold back a few students who have actually mastered it in order to avoid passing ones who have not. In other cases holding students back may be more costly than advancing students before they attain mastery.

One aspect of test design that affects the decision rules has not been covered by the previous standards, namely number of items per test or per objective. For several reasons it is important to have more than just a few items per objective. First, there must be enough items so that occasional misreading of questions by students will not often result in unwarranted failures. Second, there must be enough items so that chance effects, like guessing, do not have the opposite effect. Ideally there will be enough test questions so that three levels of attainment can be identified: clear pass, clear fail, and an area of uncertainty. Finally, a healthy number of items on a test is a protection against misjudging individual students which may result if students occasionally cooperate in answering some of the items.

Since statistics for CRTs are still largely under development, including the statistics for decision making, we are distinguishing only two general levels of merit in these rules at present.

The standard

Level C - Decision rules are not provided, or they are provided without justification, or they are hard to find or use.

Level A - Decision rules that are easy to find and use are provided along with an explanation of why they are justified.

21. Comparative or levelling data

Background

Authorities disagree on whether the intent of criterion-referenced testing is undermined by providing comparative (that is, norm-referenced) interpretations of CRTs. But test scores are not easy to interpret, and the more information that can be provided about them, the easier it is to understand them and explain them to others. Thus we are crediting CRTs that offer both absolute and relative scores with being more practical than ones that have only the former. Test users should recall that NRTs are designed to provide consistent rankings of students by selecting test items that spread out the scores of test takers. A well designed CRT (standards 1, 2, and 3) will provide less stable rankings because items are sampled to be representative of the skills or attitudes.

Note that comparative data need not be percentile norms. Average percent correct could be given for various reference groups.

The standard

Level C - Comparative data are either not provided,or are only grade level equivalents, or are based on the responses of a small or unrepresentative sample of students. Or these data are hard to find and interpret.

Level A - Acceptable comparative data are based on the responses of at least several hundred students in a nationally representative sample. Percentile norms, data for well identified reference groups, or summaries of performance of students in the target grades are suitable. These data are easy to find and interpret in the user's manuals.

REFERENCES

Campbell, D. T., Stanley, J. C. Experimental and quasi-experimental Design for Research. Chicago: Rand McNally, 1963.

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

Linn, R. L. Issues of validity in measurement for competency-based programs. Paper presented at annual meetings of National Council for Measurement in Education, New York, 1977.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.