Well-Crafted Criterion-Referenced Tests

W. James Popham

"Just as there are dull discos and yukky yogurt shops, there are criterion-referenced tests that are less fit for schools than they are for paper shredders." Six attributes of a praiseworthy criterion-referenced test are set forth in this article.

For educators, criterion-referenced tests are becoming every bit as fashionable as discotheques and frozen yogurt parlors. However, fashionable movements can attract fraudulent as well as bona fide followers. Just as there are dull discos and yukky yogurt shops, there are criterion-referenced tests that are less fit for schools than they are for paper shredders.

The Flight from Norm-Referencing

During the past decade America's educational establishment has quite properly been turning away from traditional norm-referenced tests, particularly for purposes of program evaluation and instructional diagnosis. But nature, including education, abhors a vacuum. Hence, criterion-referenced tests have recently been touted as the answer to all deserving educators' prayers. School practitioners are demanding them, and test publishers are hawking them with unprecedented zeal. It is precisely at moments like this that users of educational tests must smarten up—and in a hurry.

Whereas norm-referenced achievement tests, for several reasons beyond the scope of this discussion,[1] are decisively unsuitable for certain edu-

[1] For a one-sided analysis of the defects of norm-referenced tests for purposes of evaluation and instruction, see: W. James Popham. *Criterion-Referenced Measurement*, Chapter 4. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1978. For a more balanced appraisal of the pros and cons of norm-referenced measurement for such purposes, see: Robert L. Ebel. "The Case for Norm-Referenced Measurement." and W. James Popham. "The Case for Criterion-Referenced Measurement." *Educational Researcher*, in press. For a partisan analysis in favor of norm-referenced, scrounge for yourself.

cational purposes, badly constructed criterion-referenced tests may be little better, or even worse. To offer assistance in distinguishing between super and sordid criterion-referenced measures, a set of six attributes of a truly praiseworthy criterion-referenced test will be set forth in the remainder of this analysis. But before turning to these attributes, let's use up a few sentences attempting to clarify just what we're talking about when we use the expression, "criterion-referenced test."

The Genuine Commodity

Although there is plenty of chatter these days about just what a criterion-referenced test truly is, much of this talk is decisively on the loose side. Most educators are pretty comfortable about their understanding of what constitutes a norm-referenced test, namely, an instrument that references the examinee's performance to that of a normative group—such as when we say that "Matilda scored in the 74th percentile (in relationship to the norm group's performance)."

For criterion-referenced tests, however, there are several substantially different definitions being propounded by measurement pundits.[2] It is not the case that one of them is the *true* definition of a respectable criterion-referenced test, while all the others are imposters. No, it's just that certain conceptualizations of criterion-referenced tests offer almost no educational advantages over norm-referenced measures. And that's why educators are dumping traditional tests—they're looking for tests that will do a better educational job.

To illustrate, some people consider a criterion-referenced test to be any test for which a clearly established *level* of required examinee performance has been set. This confusion springs from the fact that for decades psychometrists have employed the term "criterion" to refer to a minimum acceptable level of examinee performance. It was only natural that this confusion would emerge when the phrase "criterion-level measurement" was originally coined.[3] But any test, even a typical norm-referenced test, can have a criterion level tagged onto it. Such a conception offers no advantages over traditional testing approaches.

Other educators think of a criterion-referenced test as one that is based on an instructional objective, typically an objective formulated in terms of desired learner behaviors. But even behaviorally-stated objectives leave too much room for interpretation; that is, such objectives can be assessed using more than one legitimate measurement procedure. Hence, the instructional objective fails to represent with sufficient clarity just what's being measured. For a criterion-referenced test, that sort of ambiguity is antithetical to the very reason that such tests were created in the first place.

What, then, is a way of conceptualizing a criterion-referenced test that offers educators the most dividends? Well, the answer is pretty straightforward. Such a criterion-referenced test is one which clearly describes an examinee's status with respect to a well-defined class of behaviors. Indeed it is the heightened descriptive

[2] An excellent analysis of the field of criterion-referenced measurement, including the matter of divergent definitions, was recently authored by: Ronald K. Hambleton, Hariharan Swarinathan, James Algina, and Douglas B. Coulson. "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments." *Review of Educational Research* 48(1): 1-47; Winter 1978.

[3] Robert Glaser. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." *American Psychologist* 18: 519-21; 1963.

power of criterion-referenced tests that makes them so superior to norm-referenced measures for purposes of instruction and program evaluation.

When educators employ a properly developed criterion-referenced test, they will know with considerable confidence just what it is that's being measured. In other words, they will know precisely what it is that the examinee can or can't do. Unlike norm-referenced measures, which only yield a depiction of an examinee's performance relative to other examinees, a good criterion-referenced test will spell out the limits of the class of behaviors being measured; for example, a competency reflecting the student's ability to comprehend the main ideas of newspaper and magazine articles.

A Well-Constructed Criterion-Referenced Test

Turning now to the qualities of a truly smashing criterion-referenced test, there are six attributes that such tests should display. Educators who are considering criterion-referenced tests for possible adoption should attend to such factors, since there are a good many so-called criterion-referenced tests that are being peddled these days that fail even to approximate these attributes. Such tests are apt to be of scant utility for most educational purposes.

An Unambiguous Descriptive Scheme

The initial, and most important, quality of a well-constructed criterion-referenced test is a descriptive scheme that with no ambiguity spells out just what it is that examinees who take the test can or can't do. Sometimes these descriptive schemes are referred to as test specifications, item forms, or amplified objectives. Whatever they are called or however they are constructed, these descriptive mechanisms are the verbal vehicles that render criterion-referenced tests useful to educators. Without crisp descriptive information that lets us know what an examinee's performance actually means, we're no better off than we were with the vague descriptions accompanying traditional norm-referenced measures.

To illustrate how test publishers can subvert the descriptive adequacy of their allegedly criterion-referenced tests, one currently popular commercially published criterion-referenced test for reading sets out its descriptions for more than two dozen objectives as follows: "When presented with a standard income tax form, the student will be able to answer multiple-choice questions about it." Such descriptive schemes are patently inadequate and potentially deceptive, since they erroneously convey the notion that test users will know what an examinee's performance actually means. All we know, of course, is that the questions based on the test form will be multiple-choice in nature. Since we have no idea of what those questions will be like, we have no real idea of what the examinee's test performance really means. Criterion-referenced tests with fuzzy descriptive schemes will not serve us well.

An Adequate Number of Items

The second quality of a spiffy criterion-referenced test: Is it assessed with an adequate number of test items? Some criterion-referenced tests attempt to assess an examinee's mastery of a given competency with only one item per competency. Other tests try to get by with as few as

two, three, or four items per measured behavior. However, it is technically impossible to get a decent fix on an examinee's status with respect to a particular skill by using only a handful of items. In situations where the stakes are high, such as when a student's graduation from high school hinges on mastering the skills represented by a test, then attempting to squeeze by with a paucity of items is both professionally and ethically irresponsible.

A Sufficiently Limited Focus

Both because we need to use enough items per measured behavior and because we don't wish to overwhelm educators with the endless lists of student behaviors so fashionable in the early days of the behavioral objectives movement, good criterion-referenced tests must be focused on a limited number of significant learned behaviors. The behaviors must be sufficiently worthwhile that they subsume a number of lesser, en route skills. At the same time, however, the small number of important behaviors being measured must still be described with sufficient clarity to communicate unambiguously what is being measured. Too many targets turn out to be no targets at all. Thus a good criterion-referenced test will focus on a limited number of skills—for example, a half dozen or so—rather than a litany-length string of 30 or more skills.

Reliability

Just as does a norm-referenced test, a well-constructed criterion-referenced test has an obligation to supply evidence that these tests possess satisfactory reliability. But there are a few new wrinkles with respect to reliability for criterion-referenced measures. For one thing, in situations where there is a high degree of learner mastery, such as might occur following very effective instruction, typical correlation-based reliability approaches will underestimate the test's reliability.

In another vein, most test users are accustomed to reliability coefficients in the neighborhood of .80 to .90. It must be recalled that such coefficients are usually based on the *total* test's reliability, and that most tests contain upwards of 50 to 100 test items. But let's imagine a criterion-referenced test that measured six separate reading skills, with 10-15 items per skill. In such cases the reliability estimates should be reported separately for the subtests, and we should not be surprised (because larger tests yield higher reliability coefficients) if the reliability coefficients for these shorter tests are markedly lower than for the longer total tests we are accustomed to using.

Validity

A good criterion-referenced test will have been subjected to a rigorous validity appraisal, particularly regarding the defensibility of the behaviors it measures. Furthermore, there should be information supplied regarding the extent to which the test's items have been demonstrated to be congruent with the test's descriptive scheme. Such evidence bears on the test's *descriptive validity.* Developers of high quality criterion-referenced tests will not overlook such validation considerations.

Comparative Data

The sixth and final characteristic of a well-constructed criterion-referenced test is interest-

ingly, the availability of normative data that will permit educators to answer more sensibly the question: "How good is good enough?" But, unlike norm-referenced tests that tie test interpretations almost completely to such normative data, a properly fashioned criterion-referenced test will retain its own descriptive clarity. The availability of comparative data is frosting on the cake, but in some instances a most necessary kind of frosting.

Detroit Public Schools' Exemplary Tests

Nowhere are the foregoing six attributes of a well-developed criterion-referenced test better exemplified than in a series of high school proficiency examinations currently being developed under the auspices of the Detroit Public Schools. Educational officials of the Detroit schools are creating criterion-referenced tests with the explicit intention of basing an improved instructional program on these measures. The tests will be the core of a minimum competency high school graduation requirement.

Three tests, each measuring four distinctive and important competencies, are being created in reading, writing, and mathematics. Each of the 12 competencies will be carefully circumscribed via sets of detailed test specifications. Equivalent forms of the tests, each containing 10 items per competency, will be employed to verify student mastery of the competencies. Many additional five-item tests will be made available to teachers for diagnostic and instructional practice. Comparative performance data will be gathered, as will evidence regarding the tests' reliability and validity.

Educators in Detroit are committed to upgrading the quality of their high schools' basic skills instructional program, and are employing properly devised criterion-referenced tests for that instructional improvement effort. A host of instructional support activities in Detroit will flow directly from the 12 clearly explicated competencies that form the basis of the new tests.

Tests That Assess, Tests That Assist

As more and more educators begin to employ properly constructed criterion-referenced tests, tests that display the six attributes described here, we will have an opportunity to see just how much instructional enhancement we may expect as a consequence of such measures. Well-devised criterion-referenced tests can supply immense instructional assistance, not merely learner assessment.

But if we fail to appraise carefully the quality of available criterion-referenced tests, naively assuming that any test that calls itself criterion-referenced must be a winner, then criterion-referenced tests are doomed to the sad fate awaiting all fashionable, but inadequately formulated endeavors. And that would be a shame. EL

W. James Popham is Professor of Education at the University of California, Los Angeles.