The Case For Norm-Referenced Measurements

ROBERT L. EBEL
Michigan State University

To build the case for norm-referenced tests and for the measurements they yield on a firm foundation, we need to begin by making clear precisely what we are talking about. In some respects, the term "norm-referenced" is unfortunate. It focuses on only one aspect, and not the most important aspect, of a large and diverse category of tests. Until Glaser referred to these tests as "norm-referenced," to distinguish them from the kinds of tests specially designed to meet the needs of programmed instruction and mastery learning which he called "criterion-referenced," they used to be referred to simply as achievement tests.

This large and diverse category includes the survey tests of spelling used by Rice in 1894, the Regent's examinations in New York State, the Iowa academic tests and tests of basic skills and of educational development, the tests developed by the Board of Examiners at the University of Chicago and at Michigan State University, and practically all published standardized tests and teacher-made classroom tests. In fact, most of the achievement tests used in the past and those currently in use belong in this category.

No sooner had the distinction between norm-referenced and criterion-referenced tests been made than we began to hear that the old-style achievement tests were really not much good, and ought to be replaced by criterion-referenced tests as quickly as possible. Had this intelligence reached the ears of the elder Thorndike, Kelley, McCall, Ruch, Brigham, Lorge and others who did so much to lay the foundations and give initial direction to the science and technology of educational measurement, no doubt they would have been surprised, and quite possibly annoyed. I will not venture to speak for our contemporaries—Linguist, Lennon, Anastasi, Cronbach, Stanley and Bob Thorndike—but if any of them has rushed to jump on the criterion-referenced bandwagon, the notice has escaped me.

Some Differences Between Norm- and Criterion-Referenced Tests

The tests called norm-referenced differ from criterion-referenced tests in a number of ways. They are considerably older, for one thing. This means that norm-referenced tests, and the technologies associated with them, have been more fully developed. Their virtues as well as their limitations are better known. The problems associated with their use have been identified and, on the whole, solved.

The two kinds of tests yield somewhat different indications of achievement. The single score that is the most important result of a norm-referenced test provides a concise summary of the examinee's general level of achievement in some area of learning. A criterion-referenced test, on the other hand, can, and often does, provide a fairly detailed and specific listing of what the examinee has or has not learned to do. This distinction can be blurred, if on a norm-referenced test the items an examinee answered correctly or incorrectly are reported, or if on a criterion-referenced test the report simply indicates how many objectives were attained. But these kinds of accommodations do not erase the essential difference.

Related to this difference in the information to be provided is a difference in the sampling of tasks or items included in the test. In a norm-referenced test, each item provides a separate and independent indication of some aspect of achievement in the domain of learning encompassed by the test. The items in a criterion-referenced test are written in clusters, each cluster focusing on a particular objective of instruction or aspect of achievement. As a result of its more extensive and more widely distributed sampling of achievements, the norm-referenced test yields a more reliable measure of a pupil's general level of achievement in a subject of study. The criterion-referenced test provides more reliable indications of whether or not some particularly important objectives have been attained.

Related also to differences in information provided and to patterns of item sampling are differences in the educational purposes to be served by the test. Norm-referenced tests are well adapted to summative evaluation, to the measurement of the results of efforts to learn. Criterion-referenced tests are more useful in formative evaluation, to guidance of the process of learning. Here again the distinction is easy to overemphasize. Deficiencies revealed by a norm-referenced test can be remedied in subsequent instruction and learning, and the number of objectives attained can be used to indicate the effectiveness of prior instruction and learning. But a difference remains in how well each kind of test can serve each educational purpose.

There is a difference in the levels of attainment that each test form is best adapted to measure. Norm-referenced tests are intended to indicate a wide range of achievement levels from outstanding excellence to serious deficiency. Criterion-referenced tests set a single standard of achievement for all. A pupil either meets or fails to meet it. To put the standard within
reach of all, it tends to be set rather low. The objectives whose attainments are to be indicated by a criterion-referenced test are often only minimums.

Proponents of criterion-referenced tests sometimes criticize the attainments tested in norm-referenced standardized tests for their lack of relevance to particular local instructional objectives. Locally designed and built criterion-referenced tests are offered as better measures than published standardized tests for the assessment of particular local instructional programs, and the success in learning of pupils in a particular local school. Of course, norm-referenced tests can be tailored to fit local instructional objectives, and criterion-referenced tests can be designed to assess the common objectives of learning in a state or a nation, but that there is some difference in emphasis between the two types of test in this respect can hardly be denied.

Finally, and most importantly, there is a difference in the conception of learning implied by the two tests. If the primary goals of learning are to acquire a series of essential abilities, distinct enough from each other, few enough in number, and important enough individually to be specified separately, studied separately, and mastered separately, then a criterion-referenced test is clearly the test that ought to be used. But if the substance of learning is an infinity of particulars, too numerous to be specified separately, studied separately, and mastered separately, then a norm-referenced test is the test that ought to be used. Such a test is now commonly referred to as a norm-referenced test.

Thus, there are important differences between norm-referenced and criterion-referenced tests, differences in the age and development of the two test forms, in the kind of information they provide, in their sampling of tasks, in the educational purposes they serve, in the range of achievement levels they measure, in the range of schools for which they are appropriate, and in the conceptions of learning they imply. But there are substantial similarities, too.

**Similarities of Norm- and Criterion-Referenced Tests**

Looked at individually, the items used in the two tests are indistinguishable. Both are intended to assess achievement in learning. The kind of tasks to be included in each can be specified precisely. The territory and the boundaries of the domain of achievements from which particular tasks are to be selected can be defined with all the precision that is necessary for either type of test.

In actual use, both kinds of tests yield scores that differ from pupil to pupil. Of course, a test of either kind might be built so as to show no score variance, but it is difficult to understand what educational value such a test would have. For since individual differences in background for learning, efforts to learn, and success in learning cannot be abolished, the only way to build a test on which all pupils would make the same score is to include tasks so simple that virtually no background for learning and no effort to learn is required for success.

Critics of norm-referenced tests sometimes claim that score variance is irrelevant to the quality of a criterion-referenced test of educational achievement. Since these critics seldom assert that differences among pupils in achievement do not exist, a proposition that would be exceedingly difficult to sustain in the face of overwhelming evidence to the contrary in all fields of human endeavor, the claim of irrelevance implies that such differences are of no concern to them. If that is so, then why test at all? The only reason to test for an achievement is that it may or may not be there. If some pupils have an achievement and others lack it, scores will vary. Score variance is not irrelevant to *any* test of achievement. Unless a criterion-referenced test of achievement is sensitive to differences in achievement, it is worthless as a test. And in general, the more sensitive a test is to such differences, the higher its quality.

Those who construct norm-referenced tests are sometimes accused of seeking score variance at any cost, of accepting it gratefully, whether or not it is related to the achievement being measured. That accusation is false. Of course, not all the variance of any test score, norm-referenced or criterion-referenced, is true score variance. There is sampling error variance. There is score variance due to bias. Test constructors are well aware of these sources of irrelevant variance. They know how to estimate their magnitudes and to minimize their effects. To do so, they pay attention to score variance, instead of claiming that it is irrelevant. Constructors of criterion-referenced tests would be well advised to follow their example.

Another point of similarity between norm- and criterion-referenced tests has to do with norms. Because criterion-referenced tests were offered in their inception as an alternative to norm-referenced tests, there was a natural tendency for the advocates of criterion-referenced tests to express disapproval of norms and of the comparisons they inevitably involved and encouraged. But when teachers and school systems began to use criterion-referenced tests, the value of norms and the need for them quickly became apparent. Norms for criterion-referenced tests are not only possible, they are essential for effective test interpretation and use. If one is interested in excellence in education, comparisons are unavoidable, for excellence can only be determined by making comparisons.

A related point, sometimes overlooked, is that norms are involved in establishing the criteria of achievement on which criterion-referenced tests are based. On what basis are standards for pupil learning set? Is it not on the basis of what pupils can learn because many of them have learned it? As Angoff has said, "Scratch a criterion and you will find a norm."

**The Case Against Easy Items**

Advocates of criterion-referenced tests criticize those who make norm-referenced tests for rejecting items that are extremely easy, particularly those that are answered correctly by all examinees. Such items, they say, are likely to test the most important outcomes of instruction. To reject them, they claim, leads to underestimation of the true achievements of pupils and schools.
The argument that the easiest items test the most important outcomes goes like this. What is most important, the school will spend the most time teaching. What it spends the most time teaching, pupils will learn the most thoroughly. If they learn it thoroughly, they will do well on items testing for it.

That is a valid argument as far as it goes, but it does not go far enough. It suggests that all pupils should learn perfectly all that their teachers try to teach. But only if the teachers were perfect teachers, and the pupils were perfect students, would it be reasonable to expect perfection in learning. Nor do we find that kind of perfection in practice. There are no 100% success items in the criterion-referenced tests used in the Michigan State Assessment. These items are designed to test the achievement of minimum competency on the most important objectives of instruction. On none of them do pupils in Michigan display anything approaching 100% success.

Note also that this argument mentions only how hard teachers try to teach something. It does not consider how hard that thing may be to learn. Some of the most important things may be the most difficult to learn. As Goethe said, “All distinguished things are difficult.” Hence, there probably is no high correlation between the importance of an educational outcome and the degree of success pupils show in attaining it. That the easiest items are also the most important is most unlikely.

Granting all this, why throw out items on which all pupils succeed? The answer, I think, is that we should not, provided that the two requirements are met. The first is that the item unquestioningly tests an achievement of unquestionable importance. Experience has taught test constructors to be suspicious of any item that all examinees answer correctly. Perhaps what it tests is common knowledge, not related at all to specific instruction. Perhaps the item includes an irrelevant give-away clue. For all or most pupils in a class, a school, a state, or a nation, to master any ability appropriate for them is a most unlikely event. An item that says they have all learned it may be suspected of lying. The second requirement to be met before retaining an item that all examinees answer correctly is that the absolute value of the test score, that is the number of items answered correctly, or the number of objectives mastered, is more important than its relative value, that is its percentile rank, stanine or z-score. This requirement focuses on one of the major differences between criterion-referenced and norm-referenced tests.

The trouble with the absolute standards of the criterion-referenced test constructor is that they are absolute only in a relative sense—relative to the standards of that particular test constructor or test scorer. And as Starch, Elliott, Ruch, and many others have shown, the standards of different test constructors and test scorers are highly personal, highly subjective, and hence highly unpredictable.

Suppose constructors of a criterion-referenced test define carefully the achievements the test is intended to measure. It is still their prerogative to aim high or low, to pursue difficult or easy achievements. If they aim low enough, they may find a number of items that most pupils can answer correctly. The pupils and the school will look good, until someone compares what they can do with what pupils in other schools can do; that is, until the criterion-referenced test constructors’ absolute standards are checked against the world’s relative standards.

If pupil achievements are going to be judged ultimately in relative terms, why not judge them in relative terms immediately? And if the judgments are to be relative, items that most pupils answer correctly have little to say about relative amounts of achievement.

Some Advantages of Norm-Referenced Tests

The question before us, as I see it, is not either/or. It is which/when. As a youth, I advanced part way up the ladder of classes and merit badges in the Boy Scouts by passing some criterion-referenced tests. In Michigan, criterion-referenced tests are being used effectively to focus attention on some of the most serious deficiencies in learning of pupils in our elementary schools. There is a place, an important place, for criterion-referenced tests in the improvement of educational achievement.

But I disagree most strongly with those who contend that criterion-referenced tests are a generally superior alternative to the more familiar and widely-used tests of achievement that have come to be called norm-referenced. Here are some advantages I see in norm-referenced tests.

1. They assess pupils’ broad general level of knowledge and understanding of a subject, not their mastery of a few particulars.
2. They reflect common nationwide goals for learning, not unique local goals.
3. They assess achievements at all levels of excellence and mediocrity. They do not focus primarily on minimum essentials.
4. Because each item can test a different aspect of achievement, they provide a broader and more representative sampling of achievements.
5. They are consistent with the view that achievement in learning is a matter of more or less, not of everything (mastery) or nothing.
6. They provide a single score that concisely summarizes pupils’ general level of achievement, not an extended inventory of things learned or not learned.
7. They are primarily useful for summative, not formative evaluation. They indicate how successful the pupils’ efforts to learn have been; how successful the teachers’ efforts to foster that learning have been.
8. They imply that the primary responsibility for successful learning rests with the pupil, not with the instructional delivery system.

It seems unfortunate to me that the advocates of criterion-referenced tests have joined the advocates of no testing at all in wholesale condemnation of conventional achievement tests. For the net effect of their combined efforts, if those efforts were to succeed, would be to discredit a good and useful tool for improving the effectiveness of education, without establishing anything nearly so good in its place. I sincerely hope that good sense will prevail and that they will not succeed.
The Case For Criterion-Referenced Measurements

W. JAMES POPHAM*
University of California, Los Angeles
Instructional Objectives Exchange

Proposition One: For Purposes of instruction or evaluation, norm-referenced achievement tests are essentially worthless. This indictment extends not only to poorly constructed norm-referenced achievement tests, but also to the very best members of this well-established measurement genre. Proposition Two: Criterion-referenced achievement tests, if they are properly fashioned, can be of enormous utility to instructors and evaluators. The issues at stake in this dispute are far more profound than is sometimes evident in theoretical disagreements between academicians. Indeed, the decisions resulting from the application of educational tests are of paramount human import.

In well over half of our 50 states, students are now obliged to display minimum competence prior to receiving a high school diploma. In most of these states, this means that high school graduates must pass a test of some sort. In instances where the wrong tests are used, and incorrect decisions are made, the self-esteem of thousands of young people will be seriously damaged, perhaps irreparably. There are also scores of decisions currently being made regarding whether to retain, revise, or scrap particular educational programs. In many instances these decisions are reached, in part, on the basis of test results. If the wrong tests are used and the wrong programs are scrapped, then countless pupils will be robbed of an effective instructional program. Clearly, the issue under analysis is more than theoretical.

But before trotting out the arguments and evidence that I hope will convince you of the soundness of my two major propositions, it is necessary to engage in a bit of stage-setting. We need to define a few terms and to retrospect briefly in order to see why this debate with my good friend and esteemed colleague, Robert L. Ebel, is even necessary.

Term-Tightening

Fortunately, there’s not much disagreement about what a norm-referenced test is. They’ve been around for so long that almost everyone concurs in viewing a norm-referenced test as a measuring device which is used to ascertain an examinee’s performance in relationship to the performance of other examinees on the same measuring device.

Unfortunately, with respect to criterion-referenced tests, there’s no unanimity regarding meaning. Much of the current ambiguity shrouding the phrase, criterion-referenced test, stems from Glaser’s classic 1963 treatise. In that essay Glaser contrasted traditional norm-referenced tests with the kinds of tests needed to assess learner performance when highly effective instruction was present, for example, such as was then being produced by programmed instructional sequences.

Yet, despite its decisively positive overall impact on educational measurement practice, Glaser’s 1963 piece allowed readers to draw two substantially different conclusions regarding what he meant by the expression, criterion-referenced test. Some people picked upon the traditional psychometric meaning of the term criterion and concluded that a criterion-referenced test was one which related examinee’s scores to a well-defined cut-off score or level of performance. Glaser’s 1963 analysis included numerous phrases, e.g., “criterion levels” and “level of proficiency,” which could lead one to such an interpretation. Other readers were more attentive to Glaser’s emphasis on the fact that a criterion-referenced test “provides explicit information as to what the individual can or cannot do.” Such readers, noting Glaser’s assertion that a criterion-referenced test can describe a student’s achievement of “specific behaviors,” inferred that a criterion-referenced test is used to ascertain an individual’s status with respect to a well-defined behavioral domain.

It is the second of these two interpretations which I shall be using throughout this debate. The former interpretation, that is, referencing an examinee’s performance on tasks (no matter how poorly defined) to a level of proficiency, offers little improvement over norm-referenced measurement strategies. On the other hand, conceiving of a criterion-referenced test as a measure of a well-defined class of behaviors offers all sorts of advantages to evaluators and instructors.

Finally, throughout this analysis I shall be focusing on achievement rather than aptitude measures. Norm-referenced aptitude measures can prove most helpful to educators, for instance, in fixed quota instructional settings where one must choose among individuals on the basis of their relative likelihood of subsequent success. My current quarrel is not with such aptitude measures but, rather, with norm-referenced achievement tests and the manner in which they have been used.
What Hath WWI Wrought?

Norm-referenced testing received its biggest boost during World War I when the Army Alpha, the first widely used norm-referenced aptitude test for groups of examinees, was administered to over 1,725,000 men. In brief, the purpose of this Herculean measurement effort was to sort out individuals according to their relative intellectual aptitude in order to make military personnel decisions. This major excursion into testing by the Army was an unparalleled success. As Du Bois observed, “It was the widespread appreciation of the Army program which greatly stimulated the making and standardizing of a wide variety of new measures and accelerated their application.”

The impact of the Army’s testing program was particularly discernible in the exponential expansion of the norm-referenced achievement tests copyrighted in the U.S. after 1920. Almost without exception, these achievement tests were constructed along the lines of the mental testing model so effectively employed by the military. These post-World War I norm-referenced achievement tests were widely misapplied, since they were employed in situations where they ought not to have been used. They were employed, for example, to appraise the quality of instructional programs, a mission not consonant with the original mental testing model of providing information for making relative contrasts among individuals. Norm-referenced achievement tests have also been touted as useful diagnostic tools, that is, instruments to aid teachers in deciding which skills to emphasize for which pupils. As most teachers will tell you, however, norm-referenced achievement tests provide little if any, useful information for instructional diagnosis.

Why shouldn’t educators have been seduced into using norm-referenced achievement tests for just about every purpose under the sun? The tests came to us professionally printed and beautifully packaged, accompanied by resplendent technical manuals loaded with enough reliability and validity coefficients to secure deference from all but the psychometrically seasoned. Moreover, norm-referenced achievement tests were born in such centers of cerebral excellence as Princeton, New Jersey. And to clinch matters, norm-referenced achievement tests became known as “nationally standardized achievement tests.” To criticize them clearly bordered on latent un-Americanism.

When norm-referenced achievement tests have been used in connection with instructional design or program evaluation, it is my belief that they have typically yielded data which are either meaningless or misleading. I shall describe three defects of norm-referenced achievement tests, arguing that it is in the nature of norm-referenced achievement tests to possess liabilities which render them unsuitable for purposes of instruction or evaluation.

Deficit One:
Teaching-Testing Mismatches

One deficit of norm-referenced achievement tests flows from the profit-making imperatives of commercial test-publishing firms. In skeletal form, the argument runs as follows:

1. Commercial test publishers must sell large numbers of their tests to stay in business.
2. But America’s historic local control of education has resulted in great curricular diversity throughout the land.
3. If test publishers spelled out explicitly what their achievement tests were measuring, many educators would see that a test was not congruent with local curricular emphases, hence would not buy the test.
4. To promote wider acceptance of their tests, commercial test publishers describe what’s being measured in very general terms, thereby making the test more appealing to a larger group of would-be purchasers.
5. The result of such imprecise descriptions is that when such tests are purchased there are often mismatches between what is taught and what is tested, such mismatches frequently being unrecognized by local educators.
6. Mismatches between what is taught and what is tested yield spurious data, resulting in misleading conclusions.

Let’s spend a moment or two with the key features of this argument. First off, is it true that there is great curricular diversity in the land? Commercial test publishers often deny it, as do proponents of norm-referenced tests. Yet careful analysis of curricular patterns by DeVault, Harnischfeger, and Wiley reveal enormous disparities in the amount of emphasis given even basic subjects. Of course most school districts have curricula which include such main-line subjects as reading, mathematics, and social studies. But it is necessary to dig beneath these superficial similarities to see that, substantial differences in curricular emphases exist at the level of detail where pupils answer test items.

Yet why is it that advocates of norm-referenced tests go through such gyrations in efforts to convince us that curriculum emphases throughout America are identical? The answer is simple. The more divergence there is in local curriculum patterns, the less suitable a nationally produced achievement test can ever be.

But in spite of all these protestations, the fact remains that our historic local control of schooling has resulted in great curricular diversity. Although commercial test publishers may attempt to mask this problem by fuzzing up their descriptive information, the problem persists.

Here, for example, are some representative examples of the level of descriptive detail supplied by most publishers of norm-referenced achievement tests.

Mathematics: Includes items measuring numeration, sets, and logic; geometry; measurement; problem-solving and operations; whole numbers

Reading Comprehension: Graded reading passages with test items based on the following objectives: literal, inferential, evaluative, and word meaning in context

Language: Includes items measuring listening comprehension; grammar and syntax; spelling; study skills; punctuation and capitalization, and usage
Even Professor Ebel ends up engaging in verbal cartwheels when he tries to defend the descriptive quality of norm-referenced tests. He points out, quite properly, that a good norm-referenced test must have more than mere relative meaning, it must also have content meaning. He then drives home his point about content meaning as follows: "These meanings and understandings are seldom wholly absent when norm-referenced measures are used." (italics added)

Since content meanings are "seldom wholly absent" in norm-referenced tests, we can rest comfortably with the assurance that such explicit meanings are almost never totally not there.

Is it any wonder that given such loose, fuzzy descriptions of what these tests measure, more than a few educators assume that the national tests match their local curricula? But those assumptions may or may not be warranted. Often, they’re not. The resulting unrecognized mismatch between what’s being taught and what’s being tested can have serious, deleterious consequences.

**Deficit Two: Imprecise Instructional Targets**

A related but distinct problem arising from the generality with which standardized tests are described, is that they provide inadequate cues for instructional design. This is true with respect to the original organization of an instructional sequence as well as formative evaluation attempts to remedy sickly instructional programs.

In shorthand form, here’s how the argument goes:

1. For economic purposes, norm-referenced achievement test publishers describe their tests very generally and also attempt to measure many skills with only a few items per skill.
2. The imprecision of these descriptions offers insufficient targets for instructional designers or formative evaluators.
3. Moreover, the inadequate number of items per measured behavior precludes reliable assessment for purposes of diagnosing the nature of an individual learner’s accomplishments.

Recent large-scale investigations have confirmed the common sense notion that time-on-task is one of the most potent variables for predicting the success of an instructional program. But if the descriptive information available with a norm-referenced achievement test is at near-Rorschach level, who can tell what kinds of behaviors should be practiced? The following description of a history subtest from a widely used social studies achievement test illustrates that point. This subtest:

...emphasizes (1) the effects of man’s increasing control over the forces of nature including conflicts with natural environment, influences of technological development, and historical change and (2) the ways in which man attempts to understand and adjust to his environment including the influences of magic, superstition, mythology, religion, philosophy, and science.

Even worse, in most norm-referenced achievement tests there are insufficient numbers of items available for the skills being measured to yield any reliable estimate of an examinee’s true mastery of that skill. Who would be willing to claim, for example, that a one- or two-item test should be used for diagnostic instructional purposes? Some norm-referenced testing devotees are even more open about this point, defending with rapture the exotic statistical manipulations that can be performed to help teachers make use of one-item subtests. 10

This severe deficit of norm-referenced achievement tests is more closely associated with the needs of instructional specialists rather than educational evaluators, but certainly evaluators will find themselves in a similar quandry if they try to revive a defective instructional program on the basis of norm-referenced achievement test results.

**Deficit Three: Psychometric Tendencies to Eliminate Important Test Items**

The third deficiency of norm-referenced tests for purposes of instruction and evaluation stems from the desire of norm-referenced test developers to construct tests that result in considerable response variance among examinees, that is, considerable differences in examinees’ test performances. Without substantial response variance it is impossible to make the fine-grained comparisons among examinees which are at the core of a norm-referenced measurement model. Furthermore, if the response variance is small, there is far less likelihood of securing the high reliability coefficients which often prove instrumental in promoting satisfactory sales of such tests.

Unfortunately, this practice has resulted in another whopping deficit for norm-referenced tests if we try to use them for instruction or evaluation. Here’s how this situation comes about:

1. Developers of norm-referenced achievement tests strive to create tests that yield substantial response variance.
2. Test items that maximize response variance possess difficulty levels of .50, thus items which are answered correctly by large proportions of the examinees must be eliminated.
3. But items which large numbers of students answer correctly often deal with topics or skills that teachers thought important, hence have stressed during instruction.
4. Therefore, particularly with oft-revised norm-referenced achievement tests, items covering important topics or skills tend over time to be excised from the tests.

Now this doesn’t mean that all important items will be removed from achievement tests, but the tendency of traditional psychometric procedures to result in the elimination of important test items is clearly present. This is surely a classic instance of Catch 22 from the teacher’s perspective. The more effectively that teachers promote a particular skill, the better their students will do on the test items measuring that skill. But the better that students do on those items, the more likely the items are to be discarded from future revisions of the test. In time, is it any surprise that frequently revised achievement tests begin to resemble aptitude tests?

When items have very low or very high difficulty levels, it is impossible to make the items discriminate effectively. In such instances, what should a test constructor do? No less an authority on norm-referenced test construction than Ebel has an answer:
If the low discrimination is due to the extreme ease or extreme difficulty of the items, they should, if possible, be revised to make them more appropriate in difficulty. If such attempts prove unsuccessful or seem certain to fail, the items should be dropped. However defensible their inclusion may be in principle, they will make little practical difference in the relative scores of the students.11 (Italics added)

Let’s be clear about what is meant by difficulty of an item. A well-instructed student should find test items easy. That doesn’t render such items intrinsically easy. It is just that after decent instruction, kids should answer test questions correctly. But the mentality of norm-referenced test developers often matches Ebel’s, when, irrespective of curricular defensibility, they discard items that fail to garner the requisite discrimination indices. This third and final deficit of norm-referenced achievement tests tends to make these measures instructionally insensitive,12 and thus valueless for both instructors and evaluators.

**Criterion-Referenced Tests as an Antidote**

Before turning to the potential advantages of criterion-referenced tests for purposes of instruction and evaluation, let’s dismiss a couple of straw person allegations that almost invariably arise when norm-referenced test proponents take out after criterion-referenced measurement strategies.13

The first of these unwarranted assumptions is that educators are opting for criterion-referenced tests instead of norm-referenced tests merely because they are reluctant to be held accountable.14 One could cogently argue that it is not the fear of accountability that terrifies educators, but the fear of inappropriately conceived accountability. Accountability schemes which center on the use of instructionally insensitive and curricularly off-target norm-referenced tests are decisively inappropriate. Educators are not turning to criterion-referenced tests because they see them as less stringent. On the contrary, criterion-referenced tests can conceivably be far more demanding than traditional norm-referenced measure. However an equitable accountability system which features the use of criterion-referenced tests will give educators at least a reasonable chance to succeed. With norm-referenced measures they have no chance at all. A second allegation which peppers the writings of those who share Professor Ebel’s dim view of criterion-referenced tests is that such tests must necessarily deal with trivial instructional outcomes, while norm-referenced tests can measure higher level outcomes. Ebel is willing, for example, to assign criterion-referenced tests to the assessment of “a limited number of specifically defined goals.”15 But when it’s necessary to tap a student’s mastery of more profound learnings, then he implies it’s time to roll out our heavy testing artillery, that is, norm-referenced tests.

But before we start dismissing criterion-referenced tests merely because they may currently deal with the puerile rather than the profound, let’s remember that in contrast to the measurement tradition of norm-referenced tests, criterion-referenced tests are still in their infancy. I readily concede that there aren’t many exemplary criterion-referenced tests around these days. That doesn’t mean that high quality criterion-referenced tests, tests that tap truly high import behaviors, cannot be created. In the embryonic field of criterion-referenced measurement, what is clearly does not reflect what can be.

**Deficit Reducing**

Earlier in this analysis three key deficits of norm-referenced tests were isolated. Let’s see now how criterion-referenced tests can take care of those problems.16

The initial deficit of norm-referenced tests arose from their inadequate descriptive power, eventuating in unrecognized mismatches between what is tested and what is taught. Since the major strength of a well-constructed criterion-referenced test is its sharpened descriptive quality, this problem is eliminated. It is eliminated, of course, assuming that the developers of the criterion-referenced test have the courage to spell out precisely what’s being measured. And it takes economic courage, since the more explicitly a testing firm spells out what’s being measured, the less likelihood that the test will be widely purchased. Many of the so-called criterion-referenced tests being peddled these days carry with them descriptive literature which is only a shade more explicit than that found with norm-referenced tests. By demanding high quality descriptions from criterion-referenced test publishers, educators will surely be able to reduce the degree of teaching-testing mismatches, and to eliminate altogether any unrecognized mismatches of that sort.

The second deficit of norm-referenced achievement tests was that they failed to provide adequate instructional targets because of imprecise descriptions and an insufficient number of items per measured behavior. Both of these deficiencies can be rectified in short order by criterion-referenced test developers. To include a substantial number of test items per measured behavior, criterion-referenced test developers must eschew the assessment of small-scale skills, attempting instead to measure truly high import competencies that subsume lesser, en route skills. By focusing on such high level skills, creating detailed descriptions of the behavior being measured, and providing an adequate number of items per measured behavior, criterion-referenced test designers can take care of deficit two with dispatch.

The third deficit of norm-referenced tests stemmed from the all-consuming need of norm-referenced test developers to produce response variance. The psychometric procedures employed to pull this off, as we saw, often resulted in the elimination of test items covering important concepts and skills. Fortunately, there is no intrinsic need to spread examinees out on criterion-referenced tests. This allows us to retain test items that many students might, after decent instruction, answer correctly. If an item is a valid index of an important skill, and many students are answering the item correctly, then we should rejoice, not despair. On all three counts, then, criterion-referenced tests can correct the inadequacies of norm-referenced tests for purposes of instruction and evaluation.

But beyond merely correcting for the deficits of norm-referenced tests...
when used for instruction or evaluation, high quality criterion-referenced tests provide a number of bonuses when used for such purposes. Most of these dividends flow directly from the greater clarity with which criterion-referenced tests spell out what they’re measuring. Moreover, this descriptive clarity does not evaporate when we measure. Moreover, this descriptive referenced tests spell out what they’re greater clarity with which criterion-referenced tests for comparative purposes. When using criterion-referenced tests accompanied by comparative performance data, we can engage in a simultaneous orgy of simultaneous cake-having and cake-eating.

As noted before, because of the recency of their introduction on the educational scene, there aren’t currently all that many outstanding criterion-referenced tests which can be wheeled out to show how really first rate they are. The technical base on which criterion-referenced tests are being constructed today is admittedly primitive.

But we are not debating the defects of poorly constructed criterion-referenced or norm-referenced tests. We are debating the adequacy of the underlying rationales associated with norm- and criterion-referenced measurement. It basically comes down to a choice between a measurement strategy which compares people versus a measurement strategy that lets us know what it is that people can or can’t do. For purposes of instruction and educational evaluation, that choice is easy.

Notes

1I am indebted to Louis Sinopoli and Joan Taylor for their assistance in assembling a variety of data needed for this analysis.

References

3Kelsey, K., & Munday, L. By the numbers. Special Report No. 2. Iowa City, Iowa: Houghton Mifflin Co., Test Department, N.D.

Let’s stop failing kids with learning disabilities.

For too long our misunderstanding of dyslexic, hyperactive and other learning-disabled children has intensified their problems. In this new book, Sylvia Farnham-Diggory shows how modern psychology has generated important advances in the understanding and treatment of learning disabilities. And shows how we can start helping these kids succeed.

Learning Disabilities
A Psychological Perspective
Sylvia Farnham-Diggory

Cloth, $7.95; paper, $3.95
From The Developing Child series

Harvard University Press
Cambridge, Massachusetts 02138

1979
Researcher’s Calendar

This 1979 calendar and reference book has ample space for notes besides each date yet puts useful general, mathematical, and statistical information close at hand. Among 11 tables are postal rates, mileages, metric unit conversions, and calories of foods. Statistical tables include a normal, chi-square, r, 5% point F and F nomograph. Several statistical procedures for correlation, dispersion, and location problems are summarized. Tukey’s quick test and the robust trimmed t are among more familiar procedures. The normal, gamma, Poisson and other densities are included along with their properties. There are integrals, series, etc. Attractively arranged, printed in two colors, the pages are sewn in signatures like fine books. It is bound with a flexible vinyl cover and is about 4 x 5 3/4 x 1/8 inches, convenient for pocket or purse. Available for $3.50 postpaid from

Dates and Bits
P.O. Box 481
Boalsburg, PA 16827
PA residents please add $.21 sales tax.