NORM–REFERENCE VS. CRITERION–REFERENCED APPROACHES TO THE MEASUREMENT OF STUDENT ACHIEVEMANT IN THE REGULAR CLASSROOM

Dr. Ezatolah Naderi
University for Teacher Education
And
Dr. Mariam Seifnaraghi
University of Alamah Tabatabai

Abstract

The use in the regular classroom of measures which are designed to yield relative indices of student achievement has been challenged severely by the recent emergence of the criterion-referenced test. The new test is designed as an absolute measure of performance. Its purpose is to assess with precision what students have achieved in relation to local instructional objectives.

In this analytical paper the authors compare the norm- referenced and criterion-referenced test in terms of purpose, construction, characteristics, and uses, and argue that classroom teachers need a new set of testing skills. Several disadvantages of the norm-referenced test for assessing student achievement as a function of instruction are emphasized.

Classroom teachers today and for always have been constructing and using achievement tests which are only relative measures. These tests tell, not **what** a student knows or can do, but **how well** the student has done in comparison with other students.

Classroom tests or standardized achievement tests which report performance in terms of the individual's relative standing in a group are called norm-referenced (N-R) tests. This is so because the test results are related to norms, which are averages obtained by a group of people who have previously taken the test and are considered representative of the new testing group.

Basic to the N-R measurement approach is a good spread of test scores, variance, from high to low so that each individual's standing in the group can be determined more reliably and so that educational decisions based on **differences** in performance can be made with greater confidence. By and large the standardized test has performed this function admirably. The classroom test not so admirably.

Into this traditional situation has arrived the criterion-referenced (C-R) test. A relatively new and fascinating development, the C-R test is an absolute measure of student achievement. These tests tell, not **how well** the student has done in comparison with other students, but precisely **what** a student has learned or can do in relation to established criteria or instructional objectives. This last phase is critical, because a C-R test score is referred to or interpreted by a criterion, or an objective standard which is pre-

sumably embedded in the statement of an instructional objective.

Basic to the C-R measurement approach is the precise assessment of an instructional objective to that the individual's performance in relation to a criterion can be more reliably determined and so that educational decisions based on **instructional effectiveness** can be made with greater confidence The need for precision in measurement in turn requires that instructional objectives be stated in clear and relatively specific language and in terms of expected student performance. Such objectives are considered to be "Measureable". Measureable student performance objectives, when they spell out what the student will know or do as a result of instruction, and how proficient the student's performance level must be to indicate achievement, are the best referents of the C-R test.

Because the two tests differ so markedly in purpose, construction, and use, the question arises as to which is prefereable for the assessment of studnet learning in the regular classroom? An alternative way of asking the question emphasizes the responsibility of an instructional program for effecting student achievement: which kind of test is better for evaluating the effectiveness of instruction?

The answer is surely the C-R test. That teachers should be constructing and using C-R tests in preference to N-R tests should be a central tenet of any teaching-learning model. The full force of the recommendation can be appreciated by comparing N-R and C-R tests and by examining the disadvantages of the N-R test for the assessment of classroom learning.

Comparisons Between Norm-Referenced and Criterion-Referenced Tests

The first comparison, already touched, concerns the relative and absolute, or subjective and objective, referents of the test scores. N-R tests are referenced to a group of people who have completed a specific test and, in consequence, have been placed at some point on the famous bell curve (normal curve or normal distribution of events). The bell-curve placement reflects an individual's score in relationship to individuals who have received similar scores or different ones. The relationships are generally reported as percentiles. So gathered, information merely informs where the individual stands in relation to peers, but not to the material that has constituted the test. To say that Mariam has scored at the 55th percentile means simply she scored better than 55 per cent of the students who have taken the test. No one yet knows what material Mariam knows or does not know.

C-R tests, alternatively, are centered on content, and scores reflect whether the individual has learned a limited amount of mateial sufficiently well. The individual's score is interpreted by its relationship to a criterion, usually called a proficiency level, and has no relationship to the scores of other individuals. We might say that N-R tests yield peer-relationship knowledge and C-R tests yield content-relationship knowledge.

A second comparison concerns the two test's content generality and content mix. The N-R test samples behavior from a broader, more diverse area of content and contains fewer items to measure each of the various content components. For this reason N-R tests can be considered global. The test measures comprehensive constructs such as reading comprehension, mathematical skills, analogous reasoning, or a relatively gross unit of instruction. Broad areas of content are needed so the test items can have a range of difficulty sufficient to produce the good spread of scores which distinguish individual achievers from one another. The global content of the N-R test unfortunately limits its ability to specify what students know and do not know in a framework from which learning and instruction can be carried on.

The N-R test, in consequence, is not a useful diagnostic tool. Nor is it sensitive to smaller units of instruction, but to individual differences on global constructs. Accordingly it is not a useful instructional tool.

The C-R test samples performance in relation to relatively specific statements of instructional objectives. Its test items exhibit a range of difficulty, but this range of difficulty is not so great, corresponding only to the content generality of the objective to which it is referenced. If the objective is item-equivalent, only one test item is needed. If the objective is content general, as is preferred, the C-R test which assesses it will have a matching range of content generality. A lengthy C-R test may contain items referenced to many instructional objectives, but such a measure yields a score for each objective, not a total test score.

Moreover, the instructional objectives from which the C-R test is derived are local. They are focused on the learning needs of a class, a school, or a district. The objectives may be specific to an individual student . Only instructional objective which most people would agree constitute the "minimal competencies" for which most programs of instruction should be responsible would be applicable to local needs as widespread as those of a country, region, or state. Because the C-R test measures local needs, it tends to be teacher-made and sensitive to local instruction. The test assesses the effectiveness of a particular teaching-learning

situation.

So, while N-R tests are sensitive to individual differences, C-R tests are sensitive to instruction. While N-R test are related to national, or statewide needs, C-R tests are related to the local needs of classes, schools, and Districts.

The next differences between the two approaches relate to methods of constructin and statistical characteristics. As mentioned, in N-R testing variance is critical, in order to discriminate between achievers and non-achievers, high achievers and low achievers. Variance establishes the statistical relationships of scores around the mean and in the situation referred to as normality, produces the familiar bell-shapped curve. Lacking variance, the discriminate power of a test is zero. Teachers would not be able to separate the good students from the bad students, an insinuating way to put it. Very importantly, variance also provides the statistical indices used by measurement experts to compute the common indicators of test usefullness, such as reliability, validity, internal consistency, and item power.

For these reasons the developers of N-R tests prefer as much variation around the mean as possible. During the tryout of items those which are answered either correctly or incorrectly by all or by too many of the respondents are eliminated from the item pool-they do not contribute to score variance. One result of the item eliminations is that areas of content many be eliminated from the test, however basic or relevant to instruction. In consequence, the N-R test contains items which approximately 50 per cent of the respondents will answer correctly and 50 per cent will answer incorrectly.

In C-R testing the factor known as precision is critical, in order to determine whether the individual has accomplished the instructional objective or met the criterion. For the test developers the important C-R test characteristic is content validity, the ability of the test to sample adequately the content of the instructional objective. Usually, several items will be required to measure a limited area of content. A high degree of variance, even some variation around the mean, is not required. The opposite is preferred, Because the test measures relatively specific objectives and is sensitive to instruction, a reduction in score variance is both expected and desirable. In C-R testing a high degree of variability can be interpreted as a function of imprecision, usually because instructional objectives are stated too vaguely or too broadly.

It is not ideal to suggest that C-R tests should be completed perfectly or nearly perfectly by large percentage of the respondents. This is because the ideal holds that C-R tests are also instructional devices, and forms of the test can be repeated during instruction until a pre-established percentage of the learners, say 85 per cent, master the designated objectives. In all cases the goal is to reduce variance, to increase mastery.

Because its chief characteristic is precision and because validity is so essential to measurement, the content validity of the C-R test must be established beyond doubt. Unlike other forms of validity, which are statistical, content validity is judgemental. It is established by the opinion of content experts. An opinion from a minimum of two experts is needed to establish reliability. The validity checks should be made by subject matter experts who did **not** take part in the development or formulation of the instructional objectives.

The content validity questions are two: (1) Does each item match the content of the instructional objective, and (2) Does the pool of items assess the full range of difficulty of the objective (obtain a representative sample of student behavior in relation to the objective?). If an item's content validity cannot be established through expert testimony, the item must be eliminated from the pool. If this continues indefinitely, the test developer can presume that something is wrong with the instructional objective and the matter becomes a curriculum development problem.

A pleasant distinction between the two tests for teachers and other practitioners is their differing reliance on statistics. N-R tests yield continuous data which are grouped as means, variance, and standard deviations, statistics which are tested for significance in complicated formula . C-R tests yield discrete data which are grouped as frequencies and percentages, statistics which are computed by the basic mathematical operations. No one need be scared off by the mathematics involved in C-R testing

A final comparison surrounds the test usage. The direct use of the N-R test is to discriminate between learners as to achievement and nonachievement on a continum or to rank students from top to bottom, from 100 to zero. The direct use of the C-R test is to determine whether an individual or group has met a pre-established criterion, or accomplished a designated instructional objective.

Although N-R tests are frequently used as measures of program evaluation (did a program or course of instruction work as well as intended?) increasing numbers of educators and test specialists have warned against that application(6). The prominent Jim Popham has declared flatly, "For purposes of program evaluation C-R tests are always preferable to N-R tests". We agree, because N-R tests cannot identify in any absolute sense what was learned by a student or accomplished by a program. It can provide only indirect or

inferred estimates of learning and accomplishment.

A particularly potent function of the C-R test which cannot be obtained from the N-R test is diagnosis. The precision of the C-R test permits it to monitor student progress throughout a course of instruction which is targeted towards many objectives, and to identify strengths and weaknesses of learners and programs. In the latter case, or in situations involving individualized instruction, test specialists[2] have urged teachers to pay more attention to C-R item scores than to objective scores. In this way individual disabilities or problems are more surely detected and instruction can be correspondingly targeted. Item scores of N-R tests can be similarly useful for diagnostic purposes but the context tends to be more disassociated from instruction. Total test scores of N-R tests are invariably too global for useful diagnosis.

When the testing purpose is to discriminate among respondents--to identify qualified school applicants, to select the best man for the job, to distribute scholarships or rewards, to rank a group of competitors, to tell how well a third-grader is reading in comparison with other third graders--the N-R test is highly preferable to the C-R test.

No doubt both the N-R test and the C-R test have their proper place in the conduct of the regular classroom, but the point to be hammered home to teachers is that the assessment of achievement as a function of instruction is not the proper domain of the N-R test.

Before moving on to more discussion of other disadvantages of the N-R test for classroom testing, it should be instructive to review the comparison already made between the two measurement approaches, as outlined in Figure 1.

Figure 1

A Comparison

Norm-Referenced Measurement	Criterion-Referenced Measurement
Relative Standards	Absolute Standards
Interpreted by scores of others	Interpreted by objective proficiency levels.
Content General	Content Specific
To increase variance	To reduce variance
More Reliability Oriented	More Validity Oriented
To rank students	To evaluate programs
Imprecise	Precise
Compares students with one another	Compares students with performance standards
Norms	Instructional objectives
Sensitive to individual differences	Sensitive to instruction
Culturally biased	Related to local needs
To Discriminate among learners	To assess student mastery
Global	Diagnostic
Yields peer-relationship knowledge	Yields content-relationship knowledge
Reported as percentiles	Reported as proficiency levels
Continuous Data	Discrete Data
Produces a predictable rate of failure	Produces a high degree of success

Disadvantages of the Norm-Referenced Test for Classroom Assessment

From the foregoing discussion comparing N-R and C-R tests, the following can be listed as fundamental disadvantages of using the N-R test in the regular classroom.

1. The tests are not content specific. They do not give concrete information as to what the students know or can do.
2. The tests are not sufficiently diagnostic. They do not measure progress in relation to objectives, or increments of learning in an instructional sequence.
3. The tests do not yield any absolute measure of the student's performance. Scores are interpreted in relation to other scores.
4. The tests do not contribute to progress evaluation, or to the improvement of program instruction.

5. The tests produce a guaranteed, predictable rate of failure; conversely, the tests do not allow for the possibility that all students, or even a great majority of students, can succeed.
6. The norm group may not represent the testing group. The norm group, to recall, is the group of people who previously completed the test and whose scores now are represented on the normal curve. This group is assumed to be representative of subsequent testing groups.

   This problem is particularly acute if the test is a nationally or regionally standardized achievement test which has been normed from those larger populations. The norm group in this case is diverse and may not adequately represent the class of students now being tested. In multicultural societies minorities regularly and rightly complain that standardized tests which they complete are culturally bias against them, that the norms to which their scores are referenced cannot fairly be applied to them.
7. N-R tests do not serve as useful instructional devices.

   Because the content is global or gross and the items are designed to identify how well students have achieved in relation to each other, the diagnostic ability of the test is weak and results are difficult to use in focusing, changing, or improving instruction.
8. N-R tests contribute to test anxiety.

   A dismal consequence of the traditional classroom testing procedures is that many students learn to fear and to dislike tests. This is because the tests are designed, not to help students learn, but to compare them with peers and to assign their grades.
9. N-R tests encourage interpersonal competition.

   A kind of competition which can become all-encompassing as students strive to compensate in other endeavors for their inability to reach the top ranks in their studies.

   Teachers can encourage impersonal competition, discouorage interpersonal competition, by shifting to an asessment procedure which sets up absolute standards within reach of the great majority, and so rewarding all students who surpass the standard.
10. N-R tests do not contribute to educational accountability.

    This last disadvantage is the saddest of all. In the broadest sense accountability means that teachers ought to assume the majority responsibility for the effectiveness of their programs. If students do not learn, the responsibility for not learning must be theirs. But, as we have plainly seen, the results of the N-R test offer no proof as to the effectiveness of instruction.

References

1. Bergman, J. **Understanding Educational Measurement and Evaluation.** Boston: Houghton Mifflin Co., 1981.
2. Block, J.H. "Criterion-referenced Measurements: **potential". School Review,** 1971, 69, 289-298.
3. **Cangelosi,** J.S. **Measurement and Evaluation.** Dubuque: WM.C Brown Co., 1982.
4. **Ebel, R.L.** "Criterion-referenced Measurements: Limitations". **School Review,** 1971, 69, 282-288.
5. **Hambleton,** R.K. and Novkdk, M.R. "Toward an integration of theory and method for criterion-referenced tests. " **Journal of Educational Measurement.**1973, 10, No.3.
6. Livingston, S.A. "Criterion-referenced applications of classical test theory". **Journal of Educational Measurement.** 1972, 9, No. 1.
7. Popham, W.J. **Criterion-referenced Measurement.** N.J: Prentice-Hall Co.,1978.
8. Popham, W.J. and Husek, T.R. "Implications of Criterion-referenced measurement." **Journal of Educational Measurement.** 1969, 6, No. 1.