DOCUMENT RESUME

ED 097 344 TM 003 954

AUTHOR Harsh, J. Richard
TITLE The Forest, Trees, Branches and Leaves, Revisited--Norm, Domain, Objective and Criterion-Referenced Assessments for Educational Assessment and Evaluation. AMEG Monograph No. 1.
INSTITUTION Association for Measurement and Evaluation in Guidance, Washington, D.C.; California Personnel and Guidance Association, Fullerton.
PUB DATE Feb 74
NOTE 15p.
AVAILABLE FROM California Personnel and Guidance Association, 654 East Commonwealth Avenue, Fullerton, California 92631 ($1.00)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Criterion Referenced Tests; Educational Assessment; *Educational Testing; *Norm Referenced Tests; *Standardized Tests
IDENTIFIERS *Domain Referenced Tests

ABSTRACT

It is argued that, by design, norm-referenced tests (NRT) and criterion-referenced tests (CRT) are conceived with different frames of reference. They are not totally exclusive of each other, but they do direct attention to different uses and references for information and decision making. Their combined contributions allow a more detailed and comprehensive means of assessing the outcomes of an educational program. A historical perspective is given of the two types of tests and NRTs are discussed as to sampling and purposes. Different types of tests are designed to sample different universes and norm-, objective-, and criterion-referenced tests are distinguished in aspects of design, development, use, and interpretation. Several of the nationally-normed achievement tests may exhibit characteristics of both NRTs and CRTs to a greater or lesser degree, according to how CRTs are defined. Criteria for evaluating educational programs, performance objectives, and the criteria of educational progress are discussed, as well as the feasibility of using CRTs in large-scale or national programs. (RC)

Association for Measurement and Evaluation in Guidance

Monograph Number 1

THE FOREST, TREES, BRANCHES AND LEAVES, REVISITED— NORM, DOMAIN, OBJECTIVE AND CRITERION-REFERENCED ASSESSMENTS FOR EDUCATIONAL ASSESSMENT AND EVALUATION

J. RICHARD HARSH

Director, Los Angeles Office
Educational Testing Service

February, 1974

Published by — California Personnel and Guidance Association

Association for Measurement and Evaluation in Guidance
A Division of American Personnel and Guidance Association
1607 New Hamphire Avenue, N.W., Washington, D.C. 20009

AMEG PRESIDENT:

Frank B. Womer

The University of Michigan

THE FOREST, TREES, BRANCHES AND LEAVES—REVISITED—— NORM, DOMAIN, OBJECTIVE AND CRITERION-REFERENCED ASSESSMENTS FOR EDUCATIONAL ASSESSMENT AND EVALUATION

"It is unfair to judge what our students have learned from that standardized test that doesn't measure the contents and emphases of our curriculum."

"You say that the students have mastered all those performance objectives. But, how well can they perform in other situations and with other contents than the specifics of your instruction?"

The Thesis

These comments are illustrative of the bi-polar arguments that have emerged regarding the virtues and limitations of norm-referenced and criterion-referenced tests. Which side of the argument attracts you? No matter! The purpose of this discourse is to argue that, by design, the NRT and CRT are conceived with different frames of reference. They are not totally exclusive of each other, but they do direct attention at different uses and inferences for interpretation and decision making. Moreover, we commend the notion that rather than viewing NRT and CRT as adversaries seeking victory over each other, their combined contributions allow a more detailed and comprehensive means of assessing and evaluating the outcomes of an educational program.

It is imperative that the consideration of this concept of the different but mutual contributions of CRT and NRT be based on assurance of the high quality of each. The limitations of the NRT or CRT are easily identified if the assessments are poorly or ambiguously constructed, administered and scored. Ill-conceived performance objectives spawn similar CRT items. Inappropriate or defective items destroy the accuracy and value of the NRT just as well as inadequate, biased or undefined population samples obliterate the possible usefulness of the NRT. In the title's analogy to the forest, it is inappropriate to consider the argument unless we begin with two healthy trees of equal quality, herein referred to as the CRT and the NRT.

Historical Observations

Historically, a form of CRT existed long before any NRT. The questions the tutor asked of his student in Greek or medieval civilizations were examples of the specific contents and purposes of instruction defining the content of the examination that would determine the student's achievement. Often the environment in which the tutoring occurred was used to illustrate the objective of the lesson, whether it was philosophy or science. In one region the question about the temperament of man was asked by analyzing the nearby olive trees; in another region the question was posed by an analogy drawn from the canals that were used to distribute the river waters to the cultivated fields. The particular competencies and values of the teacher and the diverse demands of the regions or mode of living in city, hamlet or rural isolation made variable definitions of what were relevant and important knowledge, skills or attitudes to be learned. The criteria for progress of the learner were defined and presented by and within the local situation. Such a procedure was

accepted and validated as meaningful and effective because the learner, after his schooling was completed, had to cope with knowledges, skills and attitudes of the local environment.

The 20th century brought, among other things unprecedented mobility, technology and urbanization. The children and youth experienced different education as they moved from region to region or from rural to urban environments. Moreover, youth was frequently educated in one context and soon moved to cope or find vocational adjustment in a different environment with different demands. During the 20th century, the standardized test was born out of the pressures to organize more effectively the manpower for the demands of World War I. The verbal components of these tests were soon attacked because of their bias for particular contents and environments assumed for the learner. These efforts by Army classification to devise norm-referenced common criteria were prompted by the observations of the variability of criteria of the individual examiner's judgments. At the same time, they demonstrated the problem of drawing inferences about the development of individuals with different experience and educational backgrounds. However, then and now, the critical element of the relation of educational background and accumulated learnings was imperfectly related to the tasks the individual would be required to cope with in his vocational and living demands.

Early in the 20th century, Pintner and other psychologists made exhaustive studies of the comparability of the mental developments (and accumulated achievements) of persons in different cultures and continents. Their attempts were continuously confounded by the inability to devise a measure that would be culture-free or culture-fair for the diversity of content, context and purpose of education in the various cultures. In short, separate assessments were required within the various cultures, to monitor the effects of education and the development of persons in each culture.

In the United States some of the early tests of achievement and mental ability were observed to produce different results in various regions. Particular item contents were singled out to demonstrate the bias of the item for or against individuals coming from different environments and with different educational emphasis. For example, one test item asked about the structure and uses of a single-tree. By the late 1930's, it was obvious that such content had meaning and emphasis in the education of the agrarian population, while it was seldom experienced or discussed in the urban and suburban environments of education. Conversely, the item that asked about the construction and uses of the escalator was readily seen as appropriate in urban education and relatively unknown in the rural.

These limited illustrations merely identify the historic concern for content appropriate to the purposes and context of the local or regional educational program. Insofar as the person was educated within a local context in which he would make his life, the measures of the outcomes could be readily designed for those specific knowledges, skills and attitudes that would be locally valued and required. However, as mobility became a way of life, education was concerned with helping individuals develop knowledge and skills that provided more common currency in any region of the United States. As students moved from one institution to another and from one region to another, there was interest in developing measures that were general surveys of the common skills and knowledges identified as important for coping with the inclusive culture of the country.

Purpose of Norm-Referenced Tests

Norm-referenced tests were designed to survey the skills and knowledges that were generally common to many or most educational programs. And by design, although it has rarely been recognized, the standardized norm-referenced test has an imperfect and incomplete congruence to any particular school program.

NRT's Are Designed as General Surveys

The construction of the national, standardized NRT was based on surveys of contents, materials and anticipated outcomes of schools in every region. Courses of study, curriculum guides, textbooks, instructional materials and educators' definitions were compiled and analyzed to identify contents with the highest common incidence. Items of these nationally standardized tests were designed as surveys of skills and knowledges *generally common* to many or most educational programs.

After the test was constructed, there was the further need to determine what rate and degree of attainment would be found in student populations throughout the country. To obtain an answer to this question, the test publisher defined a sampling process which would (as nearly as possible) proportionately represent the rural, suburban and urban schools in all regions of the nation. The tests were administered to this "national sample," and the performance summarized in a distribution of scores. The distribution of scores is then converted into one or several normative scales to facilitate the description or characterization of various degrees of success on the array of items in the test. The norms thus describe the range and relative incidence of success (usually with emphasis upon average or modal performance) of the reference population which is the particular obtained sample of many schools in many regions.

Sampling a Basic Element in Testing and Evaluation

There are few, if any, tests that are not designed to *sample* a very large array of contents. This is not a singular frailty of tests, for the individual in making an evaluation of another person's performance is required to make a judgment from the sample of observed behavior; and he cannot obtain observations or receive information concerning all behavior of the individual in every situation in which he is engaged.

Assessment and evaluation are basically restricted by the adequacy, representativeness and relevance of the data obtained. The sample may be an inaccurate representation of the characteristic, or it may be unrepresentative of the behavior at another time, in another format or situation.

A substantial amount of the concern with various types of tests and other assessments comes from the lack of understanding of what the technique is designed to sample, as well as the improper interpretations which are made from the data. There is a common tendency to make precise classifications of human behavior from assessment techniques that were not designed for such a purpose. Even with appropriate understanding of the test as a broad survey, or as a restricted documentation of a specific act, there is still the tendency to want to speak with precise certainty rather than with varying degrees of assurance. The basic necessity which requires *sampling* also clearly requires interpretation that describes probability and not finite certainty!

Different Types of Tests Are Designed to Sample Different Universes

To pursue the distinction of norm-, domain-, objective- and criterion-referenced measures in all aspects of design, development, use and interpretation would require a book of many pages. It is believed that this discourse may be shortened by using some figures to suggest the variety of purpose and design of the several types of tests. The figures will be restricted to the *nature of the samples* that are commonly used by the various techniques to gather information about student behavior following an educational experience.

Figure 1 presents the design of a norm-referenced test (for elementary grades 4, 5 and 6) which is used to survey many student populations on those elements which are judged to be "generally common" anticipated outcomes of education. The illustrated design also suggests that the survey may be used for several ages and thus not precisely or exhaustively be concerned with one age or program.

The illustration of the fourth grade reading domain (Figure 2) shows the (1) instructional materials (content and format), (2) instructional techniques, and (3) outcome objectives as reflected by the learning strategies and sequences of School A on the left side of the figure. On the right side of the figure are the generic categories of the reading universe found in consensus definitions of reading.

The lines with arrows show the typical match-mismatch of the test items of the generic categories with the specific contents of instruction in School A. However, it is alleged that the curriculum and instruction in School A are designed to attain the generic goals and objectives of the universe of reading.

The norm-referenced test *samples* some content from the four categories by content and format *generally* representative of generic consensus of what is included in reading.

The commercially developed criterion objectives and test items are shown to assess seven of the ten skills in the School A program, while three of the items are not included for emphasis.

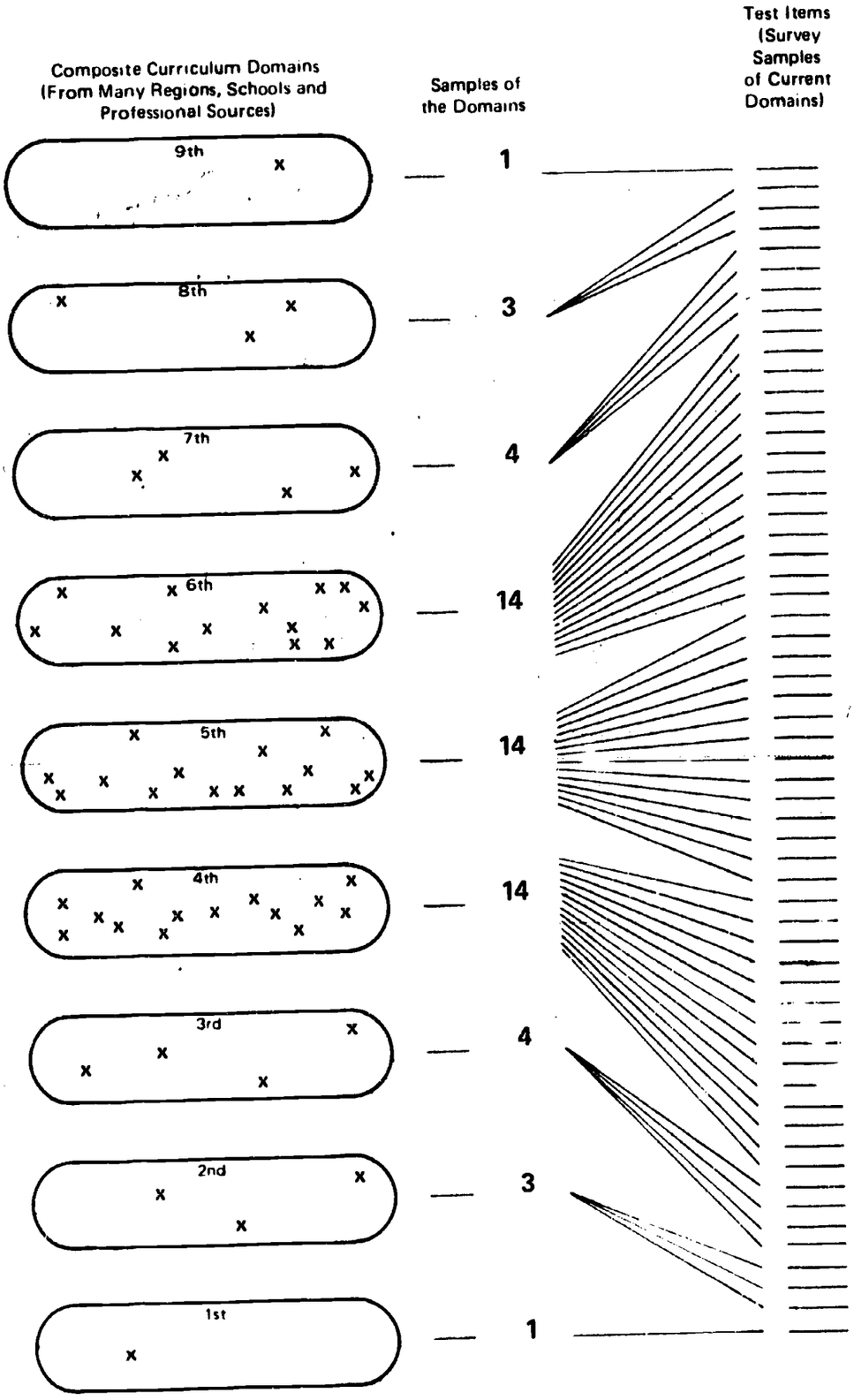
The objective-referenced test developed for grade 4 of School A provides an exact match to the objectives defined in the local reading continuum.

The criterion-referenced test for instruction in grade 4 of School A provides exact replication of the content, format and application used in the daily instructional activities. Obviously this test measures the attainment of the precise local reading experience in content and sequence.

In the schema of the 4th grade reading universe, it may be observed that the norm-referenced test and commercial domain-referenced objectives and items are designed to survey reading skills by sampling the most generic definitions of reading. The precise content, vocabulary, skills, format or application could not have a perfect match with any local program. On the other hand, the NRT provides an opportunity to survey the generalized outcomes of many different programs of reading instruction. This may be viewed as important survey information by those who exclaim, "Don't bother me with the minutia of how you teach just give me evidence that students have developed the ability to read a variety of materials they will come in contact with (beyond the materials in the daily instruction)."

Figure 1

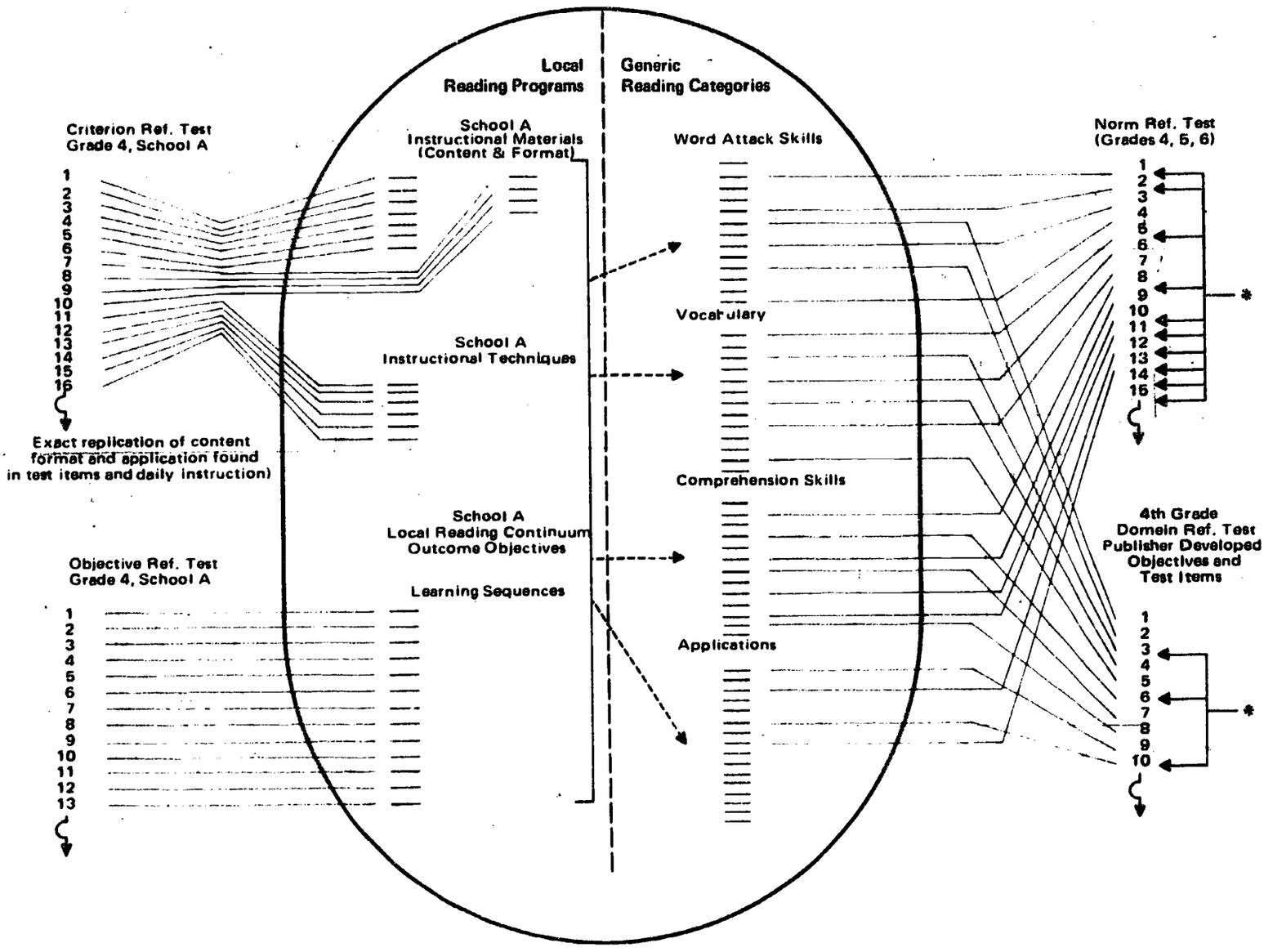
SCHEMA OF A NORM-REFERENCED TEST FOR ELEMENTARY GRADES 4, 5 AND 6



- x's represent the samples drawn from very large composite curriculum domains.
- Samples of the domains are used to develop survey test items.
- The largest samples are drawn from the 4th, 5th and 6th grade domains with very limited samples above and below the grades for intended survey testing.

Figure 2

SCHEMA OF A
FOURTH GRADE READING DOMAIN ASSESSMENT



*Publishers items have no specific match to school A 4th grade reading content, format or sequence.

It may also be observed that School A's local program has been constructed to enable students to develop the reading skills in the various generic reading categories. The dotted arrows running from the local program on the left of the schema to the generic categories indicate this planned relationship.

From the standpoint of the evaluation of the instructional program, the schema suggests that the local criterion-referenced test constructed to exactly measure the local curriculum and the local objective-referenced test should provide information of student mastery of these contents. On the other hand, the local program is said to be designed to develop student skills in the generic categories of word attack, comprehension and application. The NRT provides a general survey of these reading skills.

Different Purposes of NRT and CRT

Multiple Purposes and Uses of Testing and Evaluation

The heated discussions of the virtues and limitations of norm-referenced and criterion-referenced tests have generally ignored the different purposes and uses of these techniques and have emphasized the varying success of the student population and the congruence of the test items to the students' instructional experience. As previously stated, the NRT is designed to survey the relative attainment of students in terms of generally accepted skill and knowledge outcomes. The NRT is an external measure to provide indications of the relative achievement of many populations in relation to a reference population that is hoped to include a proportional representation of students from all types of environments and cultures of the nation.

The CRT and objective-based tests (generally of local design and construction) are planned to assess the local students' attainment of the precise content, format and sequence of their instructional program experience. The primary purpose of such measurement is to determine which specific contents and objectives have been attained and to determine the progress the students have made on the sequential objectives of the local continuums in reading, math, etc. It is generally alleged there is no interest (or appropriate procedure) to determine the relative ranking of students within or outside of the local school program. The intent is to determine mastery of locally defined performance objectives and monitor individual student progress on local curricular continuums.

In the illustration in Figure 2, the norm-referenced standardized survey test is shown to sample the universe of common reading skills. While domain-, objective, and criterion-referenced tests are given various definitions by different users, the illustration would suggest the following distinctions. A test of a domain may sample a particular sub-part of a larger universe. Objective-designed tests are usually developed to assess the particular anticipated outcomes of a local or specific instructional program. The criterion-referenced measure in this illustration is constructed to measure the mastery of the specific content, context and format of the local instructional program.

It is recognized that the foregoing distinctions are not commonly defined or exemplified by some recently developed tests given these measurement names. Certainly a portion of the lack of acceptance of these various instruments as contributing to more extensive assessment of student achievement is due to the variety of definitions and understandings of the purposes of each.

Multiple Definitions of Criterion-Referenced Tests

While the aforementioned differences in purpose and design of NRT and CRT seem basic to the issue of planning an assessment program, there are further complexities that confound the issue. Not a small problem is the multitude of ways criterion-referenced tests have been defined in the literature. The definitions are sufficiently different that a particular test may be classified as a norm-referenced test by one definition and a criterion-referenced test by another. Of even greater import is the fact that several of the nationally-normed achievement tests may exhibit characteristics of both NRT and CRT to a greater or lesser degree, according to the definition of CRT.

Hambleton and Novick have provided a thoughtful analysis of the issues and distinctions of NRT and CRT and conclude that it may be misleading to talk about NRT and CRT. They suggest that the results from either type instrument may be explained with a norm-referenced interpretation, criterion-referenced interpretation or both. What is needed is precise definition of the decision-theoretic process from which the theory, purpose and use of the measurement are derived.

Cronbach and Gleser have suggested that norm-referenced measurement is useful in situations where one is interested in a "fixed quota" selection or ranking of individuals, while criterion-referenced measurement would be useful for "quota-free" assessment. However, some recent reports of the results of criterion-referenced measurement generated per cent of students passing various items, and the users were rapidly accumulating data that might be used as a local norm. This observation reinforces the suggestion that it is time to have the *measurement theory* and the types of *uses and decisions* that are to be made from the assessment data clearly defined and understood. Then it may be more appropriate to select or develop the test that will fit the use and interpretation desired.

Criteria for Evaluating Educational Programs

A substantial basis for the argument over NRT and CRT seems to be found in the criteria that may be used to evaluate the effectiveness of educational programs. Individual schools have loudly condemned the "norms" of the norm-referenced tests for being unfair to their particular student population. The condemnation was both for the higher entry status of the modal reference population and for the norms which showed that particular school to have less than average ranking. In addition, local instructional staff became frustrated with the small increments of growth realized by the students on the normative scale as contrasted to local observations and assessments that were perceived as revealing substantial progress with the local instructional contents.

Mandatory evaluations of specially funded programs and the implementation of "accountability" procedures heightened the concern of administrators and instructional staff. The pre- post model of testing with NRT was not producing increments of growth for educationally retarded student populations equal to or greater than normally achieving student populations. These results should not have been viewed with surprise, for the entry characteristics of the retarded populations identified the lower growth increments and the additional obstacles to academic attainment that were not present in the higher-achieving populations. A thorough paradigm of learning would certainly raise a question with the assumption that any learner, irrespective of entry characteristics, would have equal opportunity for any increment of academic achievement. The accumulated past learnings

from the environment and school have been shown to be impressive determiners of subsequent learning. Unfortunately, achieving the mean or better ranking on a norm-referenced test was perceived by many as the only criterion of success.

The concern for relative ranking soon produced questions about the appropriateness of the content and format of the norm-referenced test items. Studies quickly showed that a certain per cent of the vocabulary or content of the test items was not present in a local set of instructional materials, and the students had never had practice in responding to the format of the test items. A frequent reaction was to cry, "Foul! The instrument is no good for measuring the progress of students in the local school curriculum." Needless to say, such reactions reflected a lack of understanding of the norm-referenced test as a survey that sampled generally accepted academic outcomes across a reference population of great diversity.

Local Performance Objectives as an Antidote

The creation of performance objectives stating specific contents, formats, and creditable behavior in terms of the local instructional program was viewed as a fair and exact method of determining student progress. Criterion-referenced or objective-referenced test items were then constructed to replicate exactly that specified in the local performance objectives. Review of numerous compilations of these performance objectives and their referenced test items reveals wide differences in conceptualization and technical quality. While it may be unfair to generalize grossly, it is observed that many performance objectives deal with extremely small, isolated elements of the skills of reading, math, coordination or personal-social behavior, etc.

What Are the Criteria of Educational Progress?

A review of the assessment data of several specially funded and innovative programs presents results which suggest disparate evaluations of the impact of the program on the target student population. As an illustration, the test results from an elementary school program for educationally retarded children were summarized over a three-year-period. The evaluation design called for beginning and end-of-year testing by both a NRT and a local CRT. The objectives of this program were stated as 1) 80% of the targeted students will attain mastery of 80% of the performance objectives (CRT items developed for each objective), and 2) the targeted students will attain 1.0 or more achievement on the grade equivalent scale of the norm-referenced test required by the funding agency.

The results of the first year showed the targeted population to have attained 80% or more success on 73% of the objectives. At the end of the second year, 81% of the students were reported attaining success on the objectives, and 83% attainment was shown for the third year.

During the same period in annual pre- and post-testing with the norm-referenced test, the mean change in grade equivalents was .6, .7 and .7 in the first, second and third years, respectively. The project staff observed that two years of growth had been attained in the three-year period. This was the same amount of change that had been observed in the three years prior to the project. To the project staff, the unchanging results on the norm-referenced tests were viewed as evidence that the tests were at fault and the "true" picture of growth was shown by the criterion items of the district's performance objectives.

Another result was discovered by an independent evaluator who made a longitudinal analysis of the NRT results and tested a random sample of the target students after completing three years of the program. A randomly selected group of the criterion-referenced items was used for measuring the objectives in the 1st, 2nd and 3rd years. The beginning of the 1st year to the end of the 3rd year NRT results were compared, and the difference was 1.7 on the grade equivalent scale. This was three months less than the sum of the changes observed by the pre-, post- annual testing. The CRT items also showed lower percentage mastery than had been reported in the three separate years. Of particular interest was the percentage of students showing mastery on the 1st, 2nd and 3rd year objectives. In this instance, 46% of the 1st year objectives, 53% of 2nd year objectives, and 61% of the 3rd year objectives were passed by the students in the fall semester following the completion of three years in the project (in contrast to the 73%, 81% and 83% reported at the end of the three years). These data suggest there was substantial forgetting even though the objectives dealt with academic skills that were thought to be continuously utilized in the sequential curriculum continuum.

The Nationally Standardized Test Does Not Require Norm Referencing

It is improper to draw the conclusion that nationally standardized tests must be norm-referenced; nor should it be concluded that tests designed for the assessment of local objectives and criteria are automatically freed from any ranking or norming use and interpretations. The essential issue is the need for precise definition of the design, use and interpretation for decisions that are planned for the measurement. A test item typically has a defined response which is credited as mastery of a particular element or elements of learning. This is true for very limited or comprehensive objectives of either national or local derivation.

While the majority of nationally standardized tests have been associated with norm referencing, it is quite feasible to conceptualize criterion-referenced tests in large-scale or national programs. Such a test would have items constructed to assess explicitly defined aspects of achievement, and the standardization of scoring would verify that the creditable behavior met the desired criterion of mastery. The use of such a test of objective or criterion referencing would probably be to determine whether a student or groups of students did or did not demonstrate mastery. For the individual or a group, the assessment would describe which criteria or objectives were mastered and which were not. Some psychometricians have suggested that existing standardized tests may be used in a criterion-referenced manner by identifying the items that match the desired local outcome objectives and then scoring only those items for mastery. Such usage would offer a means of assessing the mastery of designated objectives for a class, school or institution without any concern for norms or a reference population.

It is recognized that there are many technical problems involved in using either norm-referenced or criterion-referenced tests for making conclusions about the true growth of student populations. Those problems are more complex than may be adequately addressed in this paper. Suffice it to say, the reliability and validity of the measures are troublesome problems that plague those interested in very precise and parsimonious conclusions concerning short-term, annual, or longitudinal growth in academic achievement.

The issue of appropriate criteria of educational attainment is raised by these and other similar results of measurement. One cogent question relates to whether educational programs are to be judged mainly by the student mastery of local curriculum content immediately following an instructional sequence. Or is the effectiveness of the program to be viewed in relation to perseverance of accumulated learnings designed in the local program? Another cogent question is whether the purpose of many if not all local educational programs is to assist the learner in acquiring the skills, knowledges and attitudes he will need in coping with subsequent demands in school and society. Indeed, the issue is whether the educational program is concerned with the learner acquiring and internalizing the skills so that they may be applied in many formats, contexts and applications.

Many other questions may appropriately be asked about the purposes, implementation and outcomes of education for which measurement and evaluation techniques are desired. Definition of the data needed for educational decisions becomes very important to give direction to the needed measurement theory. However, prior to further clarification, it appears the criterion-referenced or objective test items designed to assess a particular instructional sequence, as well as the norm-referenced survey of more generalizable outcomes, may make mutual but different contributions of data for educational decisions.

It is common to observe the enthusiasm and effort that is mobilized for the production of a new implement or technique, even though there is inadequate definition of its purpose, use or appropriate interpretation of the results. Explicit measurement theory for the assessment of particular short- or long-term instructional experiences is a sorely needed road map. Such a map would provide direction for appropriate and effective construction and use of the norm-referenced and criterion- or objective-referenced tests.

As an epilogue, it should be mentioned that all past and present efforts to understand and assess human learning have measured only the tip of the iceberg visible above the water. A variety of techniques have been devised. At this time, new techniques and strategies are being developed. However, implements and techniques may be more effectively designed and used when there is a well-defined theory. Without the theory, potentially fine instruments may be used and interpreted with negative effects.

In quest of a theory, the road will probably lead through a forest of innovations. However, if we are to profit from the journey in measurement and evaluation, we must clearly identify the direction, terrain and destination as well as the artifacts that may be acquired along the way. Hopefully, the present ambiguity of the purpose and use of the criterion-referenced and norm-referenced tests may be seen in the perspective of the analogy of defining the forest, trees, branches and leaves of the assessment and evaluation of an educational program.

REFERENCES

Besek, R. A Comparison of Emrick and Adam's Mastery-learning Test Model with Kriewall's Criterion-referenced Test Model. Inglewood, California: Southwest Regional Laboratory, *Technical Memorandum 5-71-04*, April, 1971

Block, J. H. (Ed,) Mastery Learning: Theory and Practice. New York: Holt, Rinehart and Winston, Inc., 1971

Bormuth, J. R. On the Theory of Acheivement Test Items. Chicago: University of Chicago Press, 1970

Carroll, J. B. Problems of Measurement Related to the Concept of Learning for Mastery. *Educational Horizons*, 1970, 48, 71-80

Cronbach, L. J. & Gleser, G. C. Psychological Tests and Personnel Decisions. (2nd ed.) Urbana, Ill.: University of Illinois Press, 1965

Davis, F. B. Criterion-referenced tests: A critique. ERIC Document Reproduction Service, P. O. Drawer O, Bethesda, Md. Document Number ED 050154

Ebel, R. L. Criterion-referenced Measurements: Limitations. *School Review*, 1971, 69, 282-288

Glaser, R. & Nitko, A. J. Measurement in Learning and Instruction. In R. L. Thorndike (Ed.) *Educational Measurement*: American Council on Education, 1971, Pp. 625-670

Hambleton, R. K. & Gorth, W. P. Criterion-referenced Testing: Issues and Applications. *Center for Educational Research Technical Report No. 13* School of Education, University of Massachusetts, Amherst, 1971

Hambleton, Ronald K. and Novick, Melvin R. Toward an Integration of Theory and Method for Criterion-referenced Tests. *Journal of Educational Measurement*: NCME, Fall 1973, Pp. 159-170

Harris, M. L. & Stewart, D. M. Application of Classical Strategies to Criterion-referenced Test Construction. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971

Livingston, S. A. Criterion-referenced Applications of Classical Test Theory, *Journal of Educational Measurement*, 1972, 9, 13-26

Lord, F. M. & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968

Popham, W. J. & Husek, T. R. Implications of Criterion-referenced Measurement. *Journal of Educational Measurement*, 1969, 6, 1-9.