DOCUMENT RESUME

ED 182 324 TM 010 045

AUTHOR Mehrens, William A.; Ebel, Robert L.
TITLE Some Comments on Criterion-Referenced and Norm-Referenced Achievement Tests.
INSTITUTION National Council on Measurement in Education, East Lansing, Mich.
PUB DATE 79
NOTE 9p.
AVAILABLE FROM National Council on Measurement in Education, 1230 17th Street N.W., Washington, DC 20036 ($1.50)
JOURNAL CIT NCME Measurement in Education; v10 n1 Win 1979

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Criterion Referenced Tests; Definitions; Mastery Tests; *Norm Referenced Tests; Standardized Tests; *Test Interpretation; *Test Selection; Test Validity
IDENTIFIERS Content Validity

ABSTRACT

The discussion on criterion-referenced (CRT) and norm-referenced achievement tests (NRT) is divided into two parts: definitions and use. The authors contrast CRTs, tests which compare an individual's performance to some specified behavioral criterion of proficiency, and NRTs, tests which compare an individual's score to scores of others. They also state that any test scores must be related to test content; that NRTs also possess content validity; that allegations that standardized test publishers ignore item content in favor of item statistics are inaccurate; and that all achievement tests should be keyed to objectives or to a specified content domain. In deciding which test to use, the authors state that standardized achievement tests sample a broader domain than CRTs and are more likely to help determine the adequacy of a curriculum than CRTs which are tailor made to specific instructional objectives. CRT interpretation is described as useful for assessing mastery learning; for decision making about instructional change; and for use in broad surveys of educational accomplishment. NRT interpretation is said to be useful for rank ordering of students in specific areas of achievement and for decision making in assessing qualitative factors in addition to mastery. (MH)

Volume 10, No. 1 Winter, 1979

Some Comments on Criterion-Referenced and Norm-Referenced Achievement Tests

William A. Mehrens and Robert L. Ebel

About This Report

William A. Mehrens

Robert L. Ebel

Practitioners are today presented with a variety of tests that can be used to aid them in making instructional decisions and in curriculum planning. Some confusion has arisen concerning the distinctions among these tests, and how they can be used. The authors do much to dispel this confusion.

Drs. Mehrens and Ebel launch the first of a new series of papers which will explore the construction, interpretation and use of tests. Subsequent papers will give the readers of *ME* the opportunity to delve into such topics as applicability of latent trait theory to the *use* of tests, using standardized test results for instructional decision-making, mandated assessments and their implications for school administration, reporting educational progress to the community and selecting standardized tests for local use. Hopefully readers of *ME* will let their interests be known to this editor.

Bill Mehrens and Bob Ebel are no strangers to the readers of *ME* or to the profession of measurement. Mehrens has co-authored several prominent books including *Measurement and Evaluation in Education and Psychology*, and is a board member of NCME. Ebel, a former vice-president of ETS is the editor of the fourth edition of the *Encyclopedia of Educational Research*, has authored a new edition of his *Essentials of Educational Measurement*, and is a past-president of AERA.

HCR

The Tests: How Do We Define Them?

A Current Controversy

Measurement specialists differ in their enthusiasm for criterion referenced tests. Some see criterion referenced tests (CRTs) as the modern, improved form that all educational achievement tests should take hence forth. They feel that the problems that have plagued testing with norm-referenced tests (NRTs) will largely disappear if criterion referenced tests are substituted. Pupils will learn more, and learn it better if their efforts and those of their teachers are directed and evaluated by criterion referenced tests.

Others see a more limited, special, role for CRTs. If things to be learned are relatively few in number, separate and distinct, and if mastery of each specific ability is

possible and desireable, then criterion referenced testing is the form to use. This is likely to be true of some basic skills in the elementary grades, and in a few other areas of specialized competence. It is not likely to be true, say CRT skeptics, of most areas of study pursued in the upper grades, high schools and college.

Some Earlier Controversies

These differences of opinion over the relative merits of alternative types of tests are nothing new. They have been with us for a long, long time; probably for about as long as tests have been used. In 1845 the Boston schoolmasters resisted Horace Mann's proposal that written examinations be substituted for the prevalent oral examinations. When objective tests came into use early in the twentieth century, there were vigorous debates over their merits when compared with essay tests. Later on the focus of controversy shifted to tests of application versus tests of knowledge. In recent decades the advantages of formative tests and testing over summative tests have been argued. This paper discusses at some length the issue of criterion referenced vs. norm referenced tests.

In each of these controversies it has been apparent to all but the blindest of partisans that each type of test has its values and its limitations. Thus the real question is not "*Which* shall we use?" but "*When* shall we use it?" Stating the issue more appropriately does not guarantee an easy resolution of it, but it does improve its chances.

The Problem of Definition

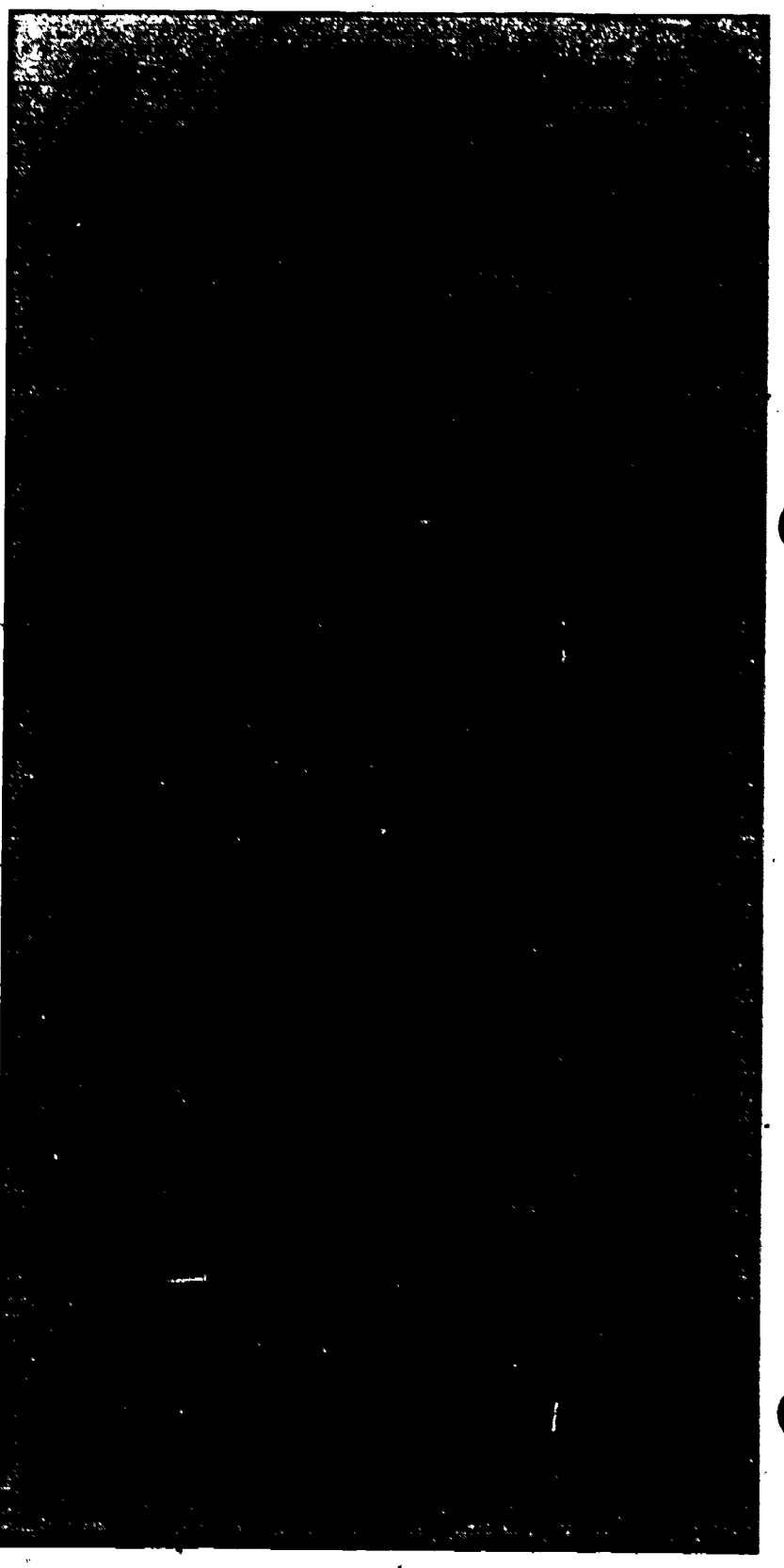
One of the difficulties in adequately addressing the question of which type of test to use is the lack of consistent definitions of some terms. Advocates of criterion-referenced tests have not always defined the concept the same way. Some of them, incorrectly we believe, make no distinction between the terms "norm referenced" and "standardized." One of the most influential advocates of criterion referenced measurement wrote:

> The key distinction, of course, between norm-referenced and criterion-referenced measurement is that in the former case we reference, that is, relate; an individual's performance to that of a norm group; in the latter case we reference an individual's performance to a criterion (Popham, 1975, p. 130).

This is clear and suggests that the distinction is between norm referenced test score *interpretation*, and criterion referenced test score *interpretation*. But the same author *defined* a criterion-referenced test as follows: "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavior domain" (Popham, 1975, p. 130).

Perhaps this kind of test is described more accurately as a "domain referenced achievement test;" as the author himself has acknowledged (Popham, 1978, p. 94).

Although there are inconsistencies with respect to how the terms norm-referenced testing (or measurement) and criterion-referenced testing (or measurement) are used, the distinction between the two types of scores seems clear enough. If we interpret a score of an individual by comparing his score to those of other individuals (called a norm group) this would be norm referencing. If we interpret a person's performance by comparing it to some specified behavioral criterion of proficiency, this would be criterion referencing. To polarize the distinction, we could say that the focus of a normative score is on how many of Johnny's peers do not perform (score) as well as he does; the focus of a criterion-referenced score is on what it is that Johnny can do. Of course we can, and often do, interpret a single test score both ways. In norm referencing we might make a statement that "John did better than 80 percent of the students in a test on addition



of whole numbers." In criterion referencing we might say that "John got 70 percent of the items correct in a test on addition of whole numbers." Usually we would add further "meaning" to this statement by stating whether or not we thought 70 percent was inadequate, minimally adequate, excellent, or whatever.

The Importance of Test Content

Although measurement experts generally agree to the distinction between norm-referenced and criterion-referenced interpretation, some misunderstanding exists about the norm-referenced interpretation. There have been those who have said, or strongly implied, that norm-referenced measurement tells us nothing about what a person can do, but only how the person compares with others, as if the comparison did not involve any specified *content* (e.g., Popham, 1976; Samuels & Edwall, 1975). It is true that the content is specified only in very general terms for some standardized tests. For many others, detailed content outlines are provided by the test publishers.

To be meaningful, any test scores must be related to test content as well as to the scores of other examinees (Ebel, 1962, p. 19). Any test will sample the content of some specified domain and there is always an implicit behavioral element. However, in norm-referenced measurement; in contrast to criterion-referenced measurement, "the inference is of the form — 'more (or less) of trait x than the mean amount in population y' — rather than some specified amount that is meaningful in isolation" (Jackson, 1970 p. 2).

Experts in achievement test construction have *always* stressed the importance of defining the specified content domain and sampling from it in some appropriate fashion. Thus, all good achievement test items, be they norm or criterion-referenced, should be keyed to a set of objectives and represent a specified content domain. If they are, the test is likely to have content validity.

Content Validity

A classic article by Lennon (1956) written *before* the current interest in CRTs, discusses three assumptions underlying the use of content validity:

1. The area of concern to the tester can be conceived as a meaningful, definable, universe of responses.
2. A sample can be drawn from this universe in some purposeful meaningful fashion.
3. The sample and sampling process can be defined with sufficient precision to enable the test user to judge how adequately performance on the sample typifies performance on the universe.

The assumptions apply to both CRTs and NRTs. One examines the content validity of any achievement test intended for a particular use by looking, among other things, at the degree to which these three assumptions are warranted. Whether or not during interpretation the *reference* is normative or criterion based is irrelevant.

Now it is quite possible that *some* locally constructed or tailor-made CRTs will have better content validity for some purposes than do standardized achievement tests which are sometimes misleadingly called norm-referenced tests. But it is wrong to imply that tests which are either standardized or norm referenced are seriously or necessarily lacking in content validity.

Content Referencing of Standardized Tests

It is also wrong to imply that only norm referencing is available for any standardized achievement test. Objective-referenced analysis can be made for scores on such tests as the *Iowa Test of Basic Skills*, *The Stanford Achievement Test*, the *Metropolitan Achievement Test* or the *Comprehensive Tests of Basic Skills*. *The Stanford Achievement Test* publisher can provide local schools with print-outs that indicate for each item (and for items grouped by instructional objectives) the behavioral objective for that item; the percent correct of that item for each class, for the total school building, for the district, and for the national standardization group. The print-outs indicate whether the local (class, building, or system) percent right is significantly lower or higher than the national percent right. They also show the response each pupil chose for each item. Obviously, the *Stanford* provides both normative and criterion-referenced data.

Item Content vs. Item Statistics

Advocates of criterion referenced tests sometimes allege that standardized test publishers are concerned almost exclusively with the statistical characteristics of their test items, and that they virtually ignore content relevance or representativeness. Consider this statement: "— *above all* [publishers of norm-referenced tests] strive to produce tests that can really spread out the norm group's performance" (italics added) (Popham, 1978, pp. 82-82). Publishers of achievement tests would deny that this is true, and a quick survey of the standard textbooks in test construction clearly shows that such striving is NOT the way to build content valid achievement tests. In discussing item analysis, Nunnally (1978, p. 264), states "it should be emphasized that item analysis of achievement tests is secondary to content validity." Ebel (1972, p. 394) and Mehrens and Lehman (1978, p. 329) make the same point. Most test publishers build tests recognizing the widely accepted priority of content validity over "good" item statistics. Certainly their test manuals typically give more coverage to content validity concerns than to item analysis approaches in item selection.

Further Definitions

It is our belief that all tests (whether CRT or NRT, whether standardized or not) should be keyed to objectives or should represent a specified content domain. Whether this process is sufficient to legitimately

allow a "criterion referenced interpretation" depends on how restrictive a definition one holds for a criterion referenced test. Ivens (1970, p. 2) simply defined a criterion-referenced test as one "comprised of items keyed to a set of behavior objectives." Harris and Stewart (1971, p. 1) gave a much more restrictive definition: "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population which the student can succeed." Glaser and Nitko (1971) define criterion-referenced tests as those deliberately constructed so as to yield scores directly interpretable in terms of specified performance standards. Millman (1974) would use the term, "domain-referenced test" for the Harris-Stewart definition of CRTs, and the term, "objective-based test" for Ivens definition.

All good achievement tests (i.e., those with high content validity) are objective based. Very few can truly be called domain-referenced (or fit the Harris-Stewart definition of a pure criterion referenced test). Most existing achievement tests probably fit in the general category of what Millman (1974) calls a criterion-referenced differential assessment device (CRDAD). In constructing such tests, one defines a content domain (but generally not with *complete* specificity) and writes items measuring this domain. But if one uses statistical procedures to judge the quality of the items with respect to their ability to differentiate groups or individuals on the degree to which they have achieved the attribute, then one lowers the confidence to be placed in an inference that a student "knows" 75% of the items correctly. The uncertainty of this particular inference is due to the use of empirical data in choosing items, whether those empirical data are pre-post test differences in item difficulty, or, are the capabilities of the items to discriminate between good and poor students at a single point in time.

Actually there are probably few situations where we need to make the pure domain-referenced interpretation. To know that an individual can type 60 words per minute is useful data whether or not the words on the tests were randomly chosen from some totally specified domain of words. To know that an individual can correctly add 80% of the items on paired three-digit whole numbers asked on a test is useful whether or not those items were randomly pulled from the total set of permutations possible.

The following distinctions among test types may bring to a focus some of the comments that have just been made.

1. *Standardized achievement tests:* These are commercial developed and may use both normative and criterion referencing. They typically sample from a broad domain of general interest and, therefore, may have less content validity for a specific purpose than a tailor-made achievement test developed just for that purpose. However, they would have more content validity for those who are interested in the broader domain.

2. *Tailor-made achievement tests:* These also may use both normative and criterion referencing. They may well be "standardized" with respect to administrative procedures. The primary distinction is that tailor-made tests are built for a specific purpose and usually sample from a constricted domain.

3. *Norm-referenced interpretation:* To add meaning to a person's score by comparing it to those of other individuals in a specified group (or groups).

4. *Criterion-referenced test interpretation:* To add meaning to a person's score by comparing it to some specified criterion of proficiency.

5. *Objective-referenced tests:* Those that are composed of tasks keyed to a set of objectives.

6. *Domain-referenced tests:* Those that consist of tasks that are sampled from a well defined population of tasks in such a fashion that one can estimate the proportion of tasks in the population at which the student can succeed.

The appropriate distinctions are between 1 and 2 on the one hand and 3 and 4 on the other, To attempt to contrast 1 and 4 (or 2 and 3) only confuses the issue. The closer any achievement test comes to fitting definition 6, the higher the content validity is likely to be. Few, if any, tests will ever have perfect content validity.

The Tests: How Do We Use Them?

Standardized vs. Tailor-Made Achievement Tests

What about the relative merits of standardized versus tailor-made achievement tests? Popham has suggested that "A growing number of educators, frustrated because the more traditional achievement tests continue to make their programs appear ineffectual, are flocking for salvation to these newer measures" (1976, p. 593). If one accepts this statement as essentially true, it raises an interesting question. Do the traditional achievement tests correctly or incorrectly show programs to be ineffectual? Are the educators turning away from traditional achievement tests because their pupils are achieving goals not tested traditionally? Is this what the supporters of criterion-referenced tests are likely to contend? Or is it because a tailor-made test avoids the possibility of comparisons between schools?

It is true, as several authors have suggested over the past few years, that standardized achievement tests have some limitations. But how much do these limitations detract from the qualities of standardized tests? How seriously do they limit the usefulness of such tests?

While there are many factors to look at in judging a test the two most important are reliability and validity. Standardized achievement tests tend to be quite reliable. The type of validity of concern is content validity. Standardized achievement tests have substantial content validity for typical school curricula. Contrary to the impressions of some critics, standardized achievement test items do measure objectives. They are based on, and sample, a specified content domain. That domain may not be specified with complete precision, and the sample may

not be completely representative. But these are limitations of every existing test, whether that test be called a standardized test, a tailor-made test, a CRT, an objectives-based test, or a domain-referenced test. No *general* statement about different degrees of content validity for standardized and nonstandardized tests is likely to be accurate. How much content validity a test has *for a particular purpose* depends on how well the items measure the objectives and sample the domain one is interested in at that time.

The charge that the items in a standardized achievement test do not match the specific objectives of a particular instructional procedure as well as an ideal test built to intentionally measure those, and only those specific objectives, is tautologically true. But does that mean we should *always* prefer the second test? Such an extreme assertion is most unlikely to be true. Consider Cronbach's words:

> In course evaluation, we need not be much concerned about making measuring instruments fit the curriculum. However startling this declaration may seem, and however contrary to the principles of evaluation for other purposes, this must be our position if we want to know what changes a course produces in the pupil. An ideal evaluation would include measures of all the types of proficiency that might reasonably be desired in the area in question, not just the selected outcomes to which this curriculum directs substantial attention. If you wish to know how well the curriculum is serving the national interest, you measure all outcomes that might be worth striving for (1963, p. 680).

Standardized achievement tests, sampling a broader domain, are more likely to help answer the question of the adequacy of the *curriculum* than are tests tailor-made to the specific instructional objectives.

The question of whether to use a more narrowly focused tailor-made test or a broader based standardized test is simply a consideration of the bandwidth/fidelity tradeoff. It seems foolish to assume that narrow bandwidth and high fidelity is always the preferred approach.

Uses for Criterion-Referenced Interpretation

The recent support for criterion-referenced interpretation seems to have originated in large part from the emphases on behavioral objectives, the individualization of instruction, the development of programmed materials, a learning theory that suggests that most anybody can learn most anything if given enough time, the increased interest in using tests for certification, and a belief that norm referencing promotes unhealthy competition and is injurious to low-scoring students' self-concepts. If we can specify important objectives in behavioral terms, then, many would argue, the important consideration is whether a student had reached those objectives, not to determine his position relative to other students.

Traditionally, the principal use of criterion-referenced measurement has been in "mastery tests." A mastery test is a particular type of criterion-referenced test. Mastery tests are used in programs of individualized instruction, such as the Individually Prescribed Instruction (IPI) program (Lindvall and Bolvin, 1967), or in the mastery learning model proposed by Bloom (1968).

Criterion-referenced interpretations are also useful in making decisions about instructional programs. In order to determine whether specific instructional treatments or procedures have been successful, it is necessary to have data about the attainment of the specific objectives the program was designed to teach. A measure that compares students to each other (norm-referenced) may not do this as effectively as a measure comparing each student's performance to the objectives.

Also, criterion-referenced measures offer certain benefits for instructional decision making within the classroom. The diagnosis of specific difficulties, followed by a prescription of certain instructional treatments, is necessary in instruction whether or not one uses a mastery approach to learning. Of course one must be very cautious about diagnosing specific difficulties on a one to five item subtest.

Finally, criterion-referenced test interpretations can be useful in broad surveys of educational accomplishment such as the National Assessment of Educational Progress or state assessment programs.

Mastery Testing

The idea of mastery learning and mastery testing is not new (see Washburne, 1922; Morrison, 1926). But the idea has not been supported unanimously. As Baker (1971, p. 65) suggested, "A considerable literature relating to the evils of mastery tests exists, and much of the work of early educational psychologists was in reaction to the unreal requirement that all pupils achieve criterion performance." The basic idea of mastery learning, however, was revitalized with the publication of a paper by Carroll (1963) entitled "A Model of School Learning." Essentially, the model suggests that the degree of learning is a function of the time the student spends on the material, divided by the time needed. More precisely, Carroll suggested that the degree of learning is some function of the time allowed and the perseverance of the student, divided by the student's aptitude for the task, his ability to understand the instruction, and the quality of instruction.

Bloom (1968) agreed with the basics of the model and suggested that the degree of learning required should be fixed at some "mastery" level and that the instructional variable should be manipulated so that all (or almost all) students achieve mastery. Bloom stated that "Most students (perhaps over 90 percent) can master what we have to teach them" (Bloom, 1968, p. 1). If the model is correct and if people should all persevere until they have "mastered" the material, then the mastery learning model of instruction should be employed and mastery testing needs to be used to determine whether mastery has occurred.

Tentative evidence (Block, 1971; 1974) suggests that in many subject-matter areas all students can achieve some level of mastery, although — as Carroll (1971, p. 31) pointed out — if the task is very difficult or depends upon special aptitudes, there may be a number of students who never make it. Becoming a four-minute miler or a concert pianist are examples of such tasks.

Excluding the extreme 5 percent of the students, the ratio between slower and faster students in the time required to master a set of objectives is about 60 to 1, although Bloom *et al* (1971, p. 51) and Bloom (1974, p. 685) have suggested that this may be reduced to about 3 to 1.

This 3 to 1 ratio is elapsed time. Bloom suggests that a more precise measure is the amount of time a student is actively working on a project. He feels the differences on this variable may be reducible to a ratio of 1.5 to 1 (Bloom, 1974, p. 688). Glaser (1968, p. 28) reported that in three years of individually prescribed instruction in mathematics, one student had covered 73 units, one only 13. Whether or not it is worthwhile educationally to have a student persist for three to six days, weeks, or years on a task that others can complete in one day, week, or year is debatable. Perhaps they should in learning those basic skills that are needed frequently by almost everyone, or that must be achieved to facilitate further learning. For other things we attempt to teach in school, such as understanding of modern literature, a mastery model should probably not be employed. There is even some doubt if it would work for such a subject. As both Block (1971, p. 66) and Bloom (1971b, p. 33) pointed out, mastery learning strategies are more effective for closed subjects (those whose content has not changed for some time) and those that emphasize convergent rather than divergent thinking. The implications for education of this admission by mastery advocates are not always fully appreciated. Cronbach brought the issue into sharp focus:

> I find the concept of mastery severely limiting and in trying to find out where my distress lies, I finally focused on one word in the Bloom paper: he states that mastery learning is closed. Training is closed. In education the problems are open . . . I see educational development as continuous and open-ended. "Mastery" seems to imply that at some point we get to the end of what is to be taught (Cronbach, 1971, pp. 52, 53).

Anastasi has made the same point. " — beyond basic skills, mastery testing is inapplicable or insufficient" (1976, p. 99).

Uses for Norm-Referenced Interpretations

Most testing and test theory has been based on the norm-referenced applications. There is little argument that such an approach is useful in aptitude testing where we wish to make differential predictions. It is also often very useful to achievement testing. Many educators would agree with Gronlund's (1971, p. 139) statement: "In measuring the extent to which pupils are achieving our course objectives, we have no absolute standard by which to determine their progress. A pupil's achievement can be regarded as high or low only by comparing it with the achievement of other pupils."

Accepting this view, the role of a measuring device is to give us as reliable a rank ordering of the pupils, with respect to the achievement we are measuring, as possible (or at least reliably place individuals into multiple categories.) Knowing what we do about individual differences, it is obvious that students will learn differing amounts of subject matter even under a mastery-learning approach. It may be that all students or at least a high percentage of them, have learned a significant enough portion of a teacher's objectives to be categorized as having "mastered" the essentials of the course or unit. But some of these students have learned more than others, and it seems worthwhile to employ measurement techniques that identify these pupils. In the first place, students want and deserve recognition for accomplishment that goes beyond the minimum. If we would continually give only mastery tests, those students who accomplish at a higher level would lose one of the important extrinsic regards of learning, that is, recognition for such accomplishments. (Of course, a CRT might not be a mastery test and might provide multiple categories. The more categories the more it discriminates like an NRT.)

Perhaps a more important reason than student recognition for discrimination testing is in its benefits for decision making. If two physicians have mastered surgery, but one has mastered it better, which one do you wish to have operate on you? For that matter, even if two physicians had equally mastered their training program, one would probably want some norm-referencing information about time to completion. If one physician is such a slow learner that it takes him five times as long as learn the material as the other one, it is probably safe to assume that after he has been on the job ten years, he will not be so up-to-date on current medical practices as the fast learner. If two teachers have mastered the basics of teaching, but one is a much better teacher, which do we want to hire? If two students have mastered first-semester algebra, but one has learned it much better (or faster, time being norm-referenced), which should receive the most encouragement to continue in mathematics? We probably all agree on the answers to these questions. However, if we have not employed measurement techniques that follow us to differentiate between the individuals, we cannot make these types of decisions. Certainly, norm-referenced measures are the most helpful in fixed-quota selection decisions. For example, if there are a limited number of openings in a pilot-training school, the school would want to select the best of the applicants — even though all may be above some "mastery level."

Excellence in any human endeavor is inescapably relative. This is as true of the learning students pursue as it is of the instruction a school provides. We cannot prevent or avoid comparisons among persons unless we are willing to give up the pursuit of excellence, unless we choose to ignore differences among people, or to defy reason by asserting that such differences are of no importance. Those who disparage norm-referenced score interpretations because they involve comparisons among persons or groups are neither soundly realistic nor beneficially idealistic.

In Conclusion

There is a place in educational measurement for both norm-referenced and criterion-referenced test interpretations. The question is not which interpretation to use, but *when* to use each. It is regrettable that we have mixed up types of tests. It is regrettable that some have advocated local tailor-made tests, not as desirable supplements to external standardized tests, which they are, but as generally superior alternatives, which they are not. Time has a way of correcting such errors. May it do so soon.

References

Anastasi, Anne. *Psychological Testing.* (4th Ed.) (New York: Macmillan Co., Inc., 1976.)

Baker, Frank B. "Computer Based Instructional Management Systems. A First Look." *Review of Educational Research*, 41: 51-70, 1971.

Block, James H. (ed.) *Mastery Learning Theory and Practice* (New York: Holt, Rinehart & Winston, Inc., 1971.)

Block, James H. (ed.). *Schools, Society, and Mastery Learning*. (New York: Holt, Rinehart & Winston, Inc., 1974.)

Bloom, Benjamin S. "Learning for Mastery." *Evaluation Comment*, UCLA, CSEIP, May 1968, 1, 2.

Bloom, Benjamin S. "Mastery Learning and Its Implications for Curriculum Development." In Elliot W. Eisner (ed.), (*Confronting Curriculum Reform*. Boston: Little, Brown, 1971.)

Bloom, Benjamin S. "Time and Learning." *American Psychologist*, 29: 682-688, 1974.

Bloom, Benjamin S., Hastings, J. Thomas, & Madaus, George F. *Handbook on Formative and Summative Evaluation of Student Learning* (New York: McGraw-Hill, 1971.)

Carroll, John B. "A Model of School Learning." *Teachers College Records*, 64: 723-733, 1963.

Carroll, John B. "Problems of Measurement Related to the Concept of Learning for Mastery." In James H. Block (ed.) *Mastery Learning Theory and Practice*, (New York: Holt, Rinehart, & Winston, Inc., 1971) p. 152.

Cronbach, Lee J. "Course Improvement Through Evaluation." *Teacher's College Record*, 64: 672-683, 1963.

Cronbach, Lee J. "Comments on Mastery Learning and Its Implications for Curriculum Development." In Elliot W. Eisner (ed.), *Confronting Curriculum Reform*. (Boston: Little, Brown, 1971) pp. 49-55.

Ebel, Robert L. "Content Standard Test Scores." *Educational and Psychological Measurement*, 22: 15-25, 1962.

Ebel, Robert L. *Essentials of Educational Measurement*. (Englewood Cliffs, NJ: Prentice Hall,) p. 394, 1972.

Glaser, Robert, "Adapting the Elementary School Curriculum to Individual Performance." *Proceedings of the 1967 Invitational Conference on Testing Problems*. (Princeton, NJ: Educational Testing Service, 1968.)

Glaser, Robert, & Nitko, Anthony J. "Measurement in Learning and Instruction." In Robert L. Thorndike (ed.), *Educational Measurement* (2nd ed.). (Washington, D.C.: American Council on Education, 1971.)

Gronlund, Norman E. *Measurement and Evaluation in Teaching* (2nd ed.) (New York: Macmillan, 1971.)

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, 1971.

Ivens, S.H. *An Investigation of items analysis, reliability and validity in relation to criterion-referenced tests*. Unpublished doctoral dissertation, 1970.

Jackson, Rex. "Developing Criterion-Referenced Tests." ERIC Clearinghouse on Tests Measurement and Evaluation, June 1970.

Lennon, Roger T. "Assumptions Underlying the Use of Content Validity." *Educational and Psychological Measurement*, 16: 294-304, 1956.

Lindvall, C.M., & Bolvin, J.O. "Programmed Instruction in the Schools: An Application of Programmed Principles in Individually Prescribed Instruction." In P. Lange (ed.), *Programmed Instruction*. 66th Yearbook, Part II. (Chicago: National Society for the Study of Education, 1967.)

Mehrens, William A., & Lehmann, Irvin J. *Measurement and Evaluation in Education and Psychology*. (2nd ed.). (New York: Holt, Rinehart, & Winston, Inc., 1978) p. 329.

Millman, Jason. "Criterion-Referenced Measurement." In Popham, W. James (ed.) *Evaluation in Education: Current Applications*. (Berkeley, CA: McCutchan Publishing Co., 1974.) Chapter 6.

Morrison, H.C. *The Practice of Teaching in the Secondary School*. (Chicago: University of Chicago Press, 1926.)

Nunnally, Jum C. *Psychometric Theory*. (2nd ed.) (New York: McGraw-Hill, 1978, 264.)

Popham, W. James. *Educational Evaluation*. (Englewood Cliffs, NJ: Prentice Hall, 1975) p. 130.

Popham, W. James. "Normative Data for Criterion-Referenced Tests?" *Phi Delta Kappan*, 57: 593-594, May 1976.

Popham, W. James. *Criterion Referenced Measurement*. (Englewood Cliffs, NJ: Prentice Hall, 1978) p. 94.

Rudman, Herbert C. "The Standardized Test Flap." *Phi Delta Kappan*, 59: 179-185 (November, 1977).

Samuels, S. Jay, & Edwall, Glenace E. "Measuring Reading Achievement: A Case for Criterion-Referenced Testing and Accountability." *Measurement in Education*. Spring 1975, Vol. 6, 2.

Washburne, Carleton W. "Educational Measurements as a Key to Individualizing Instruction and Promotions." *Journal of Educational Research*, 5: 195-206, 1922.

NCME NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

USPS 823120

1230 17th Street, N.W.
Washington, DC 20036

Second class postage paid
at East Lansing, Mich.
48823 and
additional mailing points.

A series of special reports of the National Council on Measurement in Education