An Introduction to Criterion-Referenced Measurement

P. A. Cranton
ontario institute for studies in education

L'auteur présente et définit le concept de la mesure en fonction de critères; il analyse ce qui caractérise les tests en fonction de critères à partir de quatre points: leur élaboration, leur fiabilité et leur validité, leur application et le guide qui les accompagne.

L'élaboration de ces tests comprend cinq étapes: la définition de la tâche, la composition des éléments, le choix d'un échantillon d'éléments, le choix de la longueur du test et de la note de réussite. On doit, dans le cas des tests en fonction de critères, modifier la notion que l'on a de la fiabilité et de la validité. On utilise habituellement ce genre de tests quand on n'a pas à comparer les élèves; on donne des exemples concrets. L'enseignant doit choisir entre élaborer lui-même le test ou l'acheter; l'auteur aborde cette alternative et donne quelques conseils pour l'achat de tests.

In recent years, criterion-referenced tests have appeared in most school systems in Canada and the United States, usually in conjunction with individualized instruction packages. A large literature is developing, mostly in the measurement area, in an attempt to integrate criterion-referenced measurement with existing test theory. However, the educator in the school system is often not aware of the characteristics of criterion-referenced tests and the issues that arise from these characteristics. This paper provides an introduction to criterion-referenced testing for the non-measurement oriented educator, providing him with some bases for deciding whether or not criterion-referenced tests are appropriate in a particular situation.

DEFINITION

A variety of definitions have been proposed in the literature, differing mainly in specificity. Basically, a criterion-referenced test is one which is designed to measure mastery of an objective. Unlike the general notion of "testing," students' scores are not compared one with another. Rather, a performance standard is set (for example, 3 out of 4 correct) and each individual student is compared with that standard of mastery. Other characteristics of the test are often added to the definition, but are not necessarily a part of it. The items are usually selected from a large number of possible items; most often, less than 10 items are used in one test. The task or objective must be well defined, and it must be clear from the definition which items can be included in a test. A cut-off score or a performance standard is determined prior to the administration of the test.

It is clear that a number of issues arise from these characteristics. How does one select the items and define the task? How many items should be used? What should the cut-off score be? Can an ordinary achievement test be used as if it were a criterion-referenced test? How does one know whether the test is useful or reliable? When should it be used?

TEST CONSTRUCTION

The construction of a criterion-referenced test is usually seen as involving five steps or stages: (1) defining the task, (2) developing items, (3) selecting items, (4) determining the test length, and (5) setting a passing score.

Defining the Task

Defining the task includes two processes: stating the objective or goal, and defining the items that potentially measure mastery of that objective. We will assume that the objectives are stated and concentrate on defining a group or "population" of items which could assess mastery of the objective. These two processes are very closely linked in criterion-referenced test construction since the item population is only meaningful in relation to the objective.

A number of schemes have been proposed for defining the item population (given a stated objective). One of the more common techniques is that of using "item forms." Osburn (1968) and Hively, Patterson, and Page (1968) describe this method: the structure of the item is specified and certain elements are allowed to vary, yielding a group of items measuring the same objective. For example, to assess mastery of addition of single digit numbers, the item form may be "$x + y = ?$" where "x" and "y" are less than 10. Although item forms can be relatively complex, they are generally more suited to mathematical subject areas. The advantage of this technique is that the potential items are clearly defined.

A similar but less specific method, based on "facet analysis," is described by Millman (1974). Facet design is a technique for laying out a domain for research (Guttman, 1969); facets are dimensions of a situation thought to be relevant in the measurement effort. Items are developed by combining one element from each facet. Millman suggests that a domain be defined only by those facets and elements that make a difference in how the examinee responds. To illustrate, we may have as an objective that the examinee be able to compare two objects on an equal-arm balance and choose a symbol to complete a statement of weight relation. Relevant facets may be the weight–size relationship between the two objects and the side of the balance having the heavier weight. Irrelevant facets are shapes of objects, color of the balance beam, and so on. From each relevant domain, elements are chosen and an item developed (Millman, 1974, p. 338). Facet analysis may yield vague domain definitions; however, it has the advantage of being more generally applicable to subject areas other than mathematics.

Item Development

If the task or domain has been precisely defined, item development follows easily from the definition. Hively (1970) and Hively et al. (1968) describe the procedures for generating items using the "item form" described above. A set of rules is developed for generating content within a pre-specified general framework.

When less precise domain-defining schemes are employed, item construction becomes more difficult. Millman (1974, p. 339) suggests that the test constructor should submit such items to independent reviewers in order to evaluate item–domain congruence. The judges should not have been involved in either creation of the domain or writing of the items.

Item Selection

According to many definitions of a criterion-referenced test, the items used are selected from a large pool or population of items. Theoretically, items should be selected randomly from this pool. When such a pool exists, the selection process is often performed by a computer (cf. Rosin, 1974). Since students are not being compared with each other, different items may be used for each test.

Often, however, facilities of this nature are not available. If there are more items written or seen as potential items than will be used on a test, some selection process must occur. It is not possible to use the traditional indexes of item difficulty (percentage of students passing the item) and item discrimination (degree to which the item distinguishes between low- and high-achieving students). These indexes are based on the comparison of students, which contradicts the definition of a criterion-referenced test. It is, in fact, quite likely that all students would pass a test if they had completed instruction on an objective, or fail it if it was a pre-test for an objective. Nevertheless, statistics may be used. Items which have a low correlation with the other items in the pool may not be measuring the same skill and should be dropped. Hively et al. (1968) examine the variability among items (items within "forms" should be homogeneous) using a "generalizability model." Again, poor items can be omitted.

The most practical procedure appears to be to select items from a potential or actual item pool on a random basis, taking into consideration the appropriate item statistics when they are available. This procedure helps to ensure that items are representative of the defined domain.

Test Length

In order to make accurate decisions concerning the mastery status of individual students, it is necessary to have tests of a reasonable length. Some test publishers have constructed 1- or 2-item tests to measure each skill. It is not possible to generalize from a one- or two-item test to a group of tasks or a skill. However, the teacher does not have large amounts of testing time available; a compromise needs to be found.

Millman (1974) provides tables which relate accuracy of decision-making

(precision of the test score) to test length based on a binomial model. If, for example, the test constructor can allow a 10% chance that a student whose "true" score is 90% will receive less than 70% on the test, he can use a 6-item test.

A more complex method of determining test length uses a Bayesian model (Novick & Lewis, 1974; Hambleton & Novick, 1973; Swaminathan, Hambleton, & Algina, 1975). Some tables have been prepared (Novick & Lewis, 1974) containing recommended test lengths (generally shorter than those obtained from the binomial model mentioned above). The actual calculations performed are complex: the reader is referred to the articles mentioned and also to Millman (1974) for details. The test constructor must be able to state some prior beliefs about performance and make some assumptions concerning the distribution of scores; therefore, the model may not be appropriate in all situations.

Another strategy that has been developed is to make the test length dependent on the performance of the student (Ferguson, 1970). If the student is clearly failing (2 or 3 items wrong sequentially), testing is terminated; if the student is clearly passing (4 or 5 items correct), testing also terminates; if the status is not clear, testing continues. This technique is most feasible in a computerized situation, but could be utilized in the classroom under some circumstances (e.g. with programmed booklets).

Generally, the overly brief tests should be avoided. Some useful tables have been developed to aid the test constructor. In a computer-assisted instruction setting, the educator may want to investigate the use of varying test lengths ("tailored testing").

Determining the Passing Score

A number of guidelines have been proposed for the determination of a passing or cut-off score for the criterion-referenced test. Smythe, Kibler, and Hutchings (1973) suggest that the teacher make a decision based on experience and modify it as necessary. Also, if performance data from previous samples of students are available, they may be used. It may be important that a predetermined percentage of the students pass; Tinkleman (1971) states that this could specify the passing score. Millman (1973, 1974) provides several guidelines. A passing score can be established from the past relationship between scores and the effect of the decisions made from them. The educational consequences (effect on future learning) may be considered – a high proficiency level should be required when knowledge and skills are fundamental or prerequisite to future learning. A lower passing score can be tolerated when the material does not complete a necessary link in the development of some more complex concept or skill. Psychological and/or financial cost can be considered; that is, there should be fewer failures when the "cost" (lower motivation, boredom, damage to self-concept, dollars, time) is high. Ebel (1973) suggests that the importance of individual items or groups of items might be considered. He devises

a grid of item importance by item difficulty, showing the proportion of items in each cell that must be passed by a minimally qualified person.

In a simulation study, Spineti and Hambleton (Note 1) investigated, among other factors, the effect of using a variety of cutting scores. They found that the use of an 80% cut-off score with a 5-item test yielded the fewest errors in decision-making.

It should be noted that it is not always necessary to set a passing score. If no decision as to mastery or non-mastery is to be made, no passing score is needed: the teacher may want an estimate of a student's knowledge of an objective and have no need to base a decision on this estimate. But, given that a passing score is necessary, the specification of that score is primarily subjective. The guidelines presented above should be utilized where appropriate and where the information is available to yield a more meaningful decision-making process.

RELIABILITY AND VALIDITY: TEST EVALUATION

After the criterion-referenced test has been constructed (or purchased) the educator will likely want to evaluate the test. The traditional methods of evaluating tests are to estimate reliability (or consistency) and validity (or accuracy). These notions continue to be relevant in criterion-referenced measurement, with some modifications.

Reliability

Traditional tests often employ correlational estimates of reliability; that is, how consistent are items in the test and how stable is the test over time. However, the calculation of a correlation is based on variance among the scores. A large variance will not necessarily exist among criterion-referenced test scores; correlations, then, will not give accurate reliability estimates.

A number of alternate techniques have been proposed. Carver (1970) suggests that reliability be estimated by administering the test to two similar groups and comparing the percentage that achieved the passing score in the two groups. A similar approach, which has been developed in more detail, looks at consistency of the test scores across repeated administrations (Swaminathan et al., 1974). The authors describe a coefficient which can be calculated by the test constructor as a measure of agreement between the different occasions. In a third technique Millman (1974) measures the consistency of scores for students who are given two sets of items drawn from the same domain (parallel tests). The two tests are given on the same occasion; in fact, it is suggested that the items be intermingled. The discrepancy between the two scores can be examined, or the agreement of the decisions made from the scores can be computed (i.e. the proportion of students placed into the same decision category by both tests).

Validity

If a test is valid, the score from that test gives an accurate picture of the student's knowledge of the area the test was designed to measure. Validity,

then, is closely related to the test construction procedures; if items were developed in accordance with the definition of the domain, the test is more likely to be valid. It is suggested by some authors that validity can best be determined by a careful and logical analysis of the domain definition, the item construction procedures, and the apparent relevance of individual items (Popham & Husek, 1969; Millman, 1974).

If the test constructor is uncomfortable with these subjective analyses, some alternatives are available. If item forms are utilized in test construction, the set of items produced from an item form can be examined for homogeneity, using a generalizability model (Hively et al., 1968) or an index of homogeneity (cf. Macready & Merwin, 1973). The statistics will give some indication that the items are really related to the defined domain. A second procedure may be to make some predictions about the test, then to check these predictions. For example, the test items may be related to objectives that are part of a validated hierarchy of skills. Theoretically, a student should not pass items measuring an objective higher in the hierarchy nor fail items measuring an objective lower in the hierarchy. The test validity may be determined, then, by the extent to which the test scores substantiate the hierarchy. Millman (1974, p. 361) describes a coefficient which can be used in this situation.

APPLICATIONS

The uses of criterion-referenced measurement are almost implicit in the definition. However, there is often confusion, due in part to the tendency of educators to "jump on the bandwagon" to use the latest technique.

Generally, criterion-referenced testing is a technique used when there is no need to rank or compare students — when the information required concerns only knowledge of an objective or mastery of a skill.

The teacher who is planning an instructional unit may need to know the students' level of knowledge on a number of objectives. For example, if one plans to teach multiplication, it is important to check addition skills. Or, in a series of non-hierarchical objectives, the students may have knowledge of some and not others; a set of criterion-referenced tests provides this information for instructional planning. Students need not be compared; it may not even be necessary to make mastery/non-mastery decisions.

If a student is learning a set of hierarchical skills, it is necessary to determine whether one skill is mastered before the next is attempted. In the study of basic skills, it is especially important that mastery decisions be made accurately. Here the criterion-referenced test, having items constructed from skill definition, is most likely to yield accurate decisions.

It is sometimes necessary to judge the effectiveness of an instructional sequence, whether it be a programmed booklet, a computer-assisted instruction program, or a segment of classroom teaching. If two parallel forms of a criterion-referenced test are given, one before and one after the instruction, the difference between these scores gives some indication of

the effectiveness of the intervening instruction – that is, the degree to which the objectives have been met. Millman (1974, p. 392) also suggests that teacher performance could be evaluated using this information; however, numerous other variables are involved in the teacher–student interaction.

In an individualized instruction package (either paper and pencil or computerized) it is necessary to manage or at least monitor the progress of the student in some way. The criterion-referenced test is often used in this way even if the objectives are not hierarchical. Many packages require mastery of each set of objectives before the student can proceed. Often remedial sequences are available for students in the non-mastery status. Serious difficulties can be noted by reviewing test scores regularly.

Why Not Use "Ordinary" Achievement Tests for These Purposes? Traditional tests, such as the *Canadian Tests of Basic Skills* or the *Gates MacGinitie Reading Test,* are called norm-referenced tests – students are compared to a "norm" or an average and are ranked according to their scores. The question of using these tests to perform the function described for a criterion-referenced test is often raised. Generally, the items on a norm-referenced test are chosen for their discriminating ability – that is, for how accurately they rank or categorize students. Such items are not the same as items selected randomly from a domain: the components of the domain which students are likely to have in common are eliminated. Criterion-referenced test items are constructed from a (usually) precisely defined domain and validated according to the accuracy with which they measure knowledge of that domain. Norm-referenced test items may or may not be constructed from a defined domain; items are selected on the basis of item statistics, and high reliability is dependent on large variation among scores. One should not consider using a norm-referenced test for criterion-referenced measurement: it is not necessarily generalizable to the task, and it likely measures a more general ability or a subset of the task that discriminates among individuals. Conversely, one would rarely consider using a criterion-referenced test for a norm-referenced function: it has no norms, no basis for comparing scores, and unknown discriminating ability.

AN EVALUATION AND GUIDELINES FOR USERS

Numerous articles have appeared discussing the advantages and disadvantages of criterion-referenced testing. Unfortunately, the majority do not seem to consider the purposes for which the different tests were intended. A small sample of this literature will be examined.

Brazziel (1972) lists several advantages and disadvantages of criterion-referenced testing. The benefits that he sees include the interpretation of progress in terms of objectives, the facilitation of individualized instruction, the elimination of the concept of one-half the students being below the mean, the shortness of the tests permitting regular measurement,

the elimination of "teach to the test" pressures, and the possibility of compiling comprehensive records on child development. On the other hand, he sees criterion-referenced tests as hindering reporting systems for students who move from one school to another, lacking construct validity, hindering the comparison of school districts, and requiring excessive teaching materials for specific objectives. Some of these points are not valid. First, although one-half of the students are no longer below the median, there are still a certain number who are "non-masters" and must repeat instruction or receive remedial instruction. Surely this is as damaging to the child's self-concept as being below the median. Secondly "teach to the test" may better be emphasized than eliminated: the items are directly related to objectives which have to be explicitly taught in order to maximize the student's chances of passing the test. Third, if one is interested in comparing school districts, it is likely possible to administer a norm-referenced test for this purpose. Finally, it is not likely that a teacher would have available a set of criterion-referenced tests and have no teaching materials for those objectives.

Ebel (1971) discusses limitations of criterion-referenced measurement. He states that such tests do not tell us all or even the most important part of what we need to know about educational achievement. Good criterion-referenced measures are difficult to obtain – they require a degree of detail in the specification of objectives or outcomes that is unrealistic to expect and impractical to use. Also, Ebel argues that mastery should not necessarily be a teaching goal: abilities, understandings, and appreciations are not all-or-none adaptations. Block (1971) refutes all of Ebel's statements. Block claims that much school learning is sequential and that criterion-referenced tests tell us specifically what the student has not learned at a certain stage. In answer to Ebel's second point, Block says that objectives stated in enough detail to guide the teaching–learning process are also adequate to develop criterion-referenced tests. Also, published objectives and tests are available and computer-generated tests may be feasible in some cases. Third, Block argues that skills requiring mastery do form a large portion of education and are essential.

A final illustration of this type of debate is an article by Shoemaker (1971). Shoemaker claims that a teacher needs to know the relation of each individual to the group, and that this information is not supplied by the criterion-referenced test. Scores on a criterion-referenced test are skewed and therefore provide minimal information for a classroom instructional management system. Other general points are made about the characteristics of criterion-referenced measurement. In answer to such arguments, it can only be said that if comparisons are to be made and if normally distributed scores are desired, criterion-referenced tests should not be used.

Generally, then, the question of the advantages and disadvantages of criterion-referenced measurement is related to the purpose for which the tests were intended. Uses for criterion-referenced tests have been outlined

above. Millman (1974, p. 319-323) provides examples to illustrate the situations in which different types of tests are appropriate. If the educator wishes to assess a student's knowledge of an objective, or mastery of a skill relative to some standard, then a criterion-referenced test is a useful instrument. If, however, the educator wants to compare students in any way, to divide them into groups, or discriminate among them, a criterion-referenced test is the wrong tool.

Writing or Buying Tests

If it is decided that criterion-referenced tests are desirable in a certain educational setting, two choices are available – developing or buying tests.

Following the recommended procedures for constructing criterion-referenced tests is time-consuming. Objectives must be clearly stated, items developed directly from the objectives, items selected, and the reliability and validity of the tests checked. However, if enough time is available or if objectives are already in existence, the task may be worthwhile. Tests that are developed locally have the advantage of being directly relevant to the goals of the teacher, to the particular community and school situation, and to the conditions surrounding certain groups of students. The usual problem of American flags and holidays in Canadian classrooms and "downtown" concepts in rural settings is eliminated.

However, purchasing published tests is obviously more convenient. If this is the decision, a few precautions should be taken. The lengths of the tests should be checked with one of the tables relating test length to accuracy (mentioned in the section on test length). Tests of one or two items should never be accepted as adequate if any decision is to be made on the basis of the test results. Descriptions of test construction procedures and test evaluation figures should be available and should be examined carefully. Without this information, there is no possibility of generalizing from the test scores to any wider domain. Finally, read the test material carefully and determine if it is actually relevant to your curriculum and your situation. If not, can it be modified easily? Would the modifications affect the reliability and validity of the test?

CONCLUSION

Criterion-referenced tests can yield valuable information about the performance of individual students. If they are used in appropriate situations, and if the tests themselves are carefully evaluated, the teacher and students are likely to benefit from the information they yield. Criterion-referenced measurement is currently enjoying widespread popularity in the schools, and theoretical literature is flourishing. However it is often the case that the persons using the tests do not come into contact with or are uninterested in the theoretical articles. It is hoped that this paper is one step in bridging that gap.

NOTE

P. A. Cranton is currently affiliated with McGill University.

REFERENCE NOTE

1. Spineti, J. P., & Hambleton, R. K. *A computer simulation study of tailored testing strategies for objective-based programs.* Unpublished paper, 1974.

REFERENCES

Block, J. H. Criterion-referenced measurement: Potential. *School Review*, 1971, *79*, 289–298.

Brazziel, W. Criterion-referenced tests: Some trends and prospects. *Today's Education*, 1972, *61*, 52–53.

Carver, R. P. Special problems in measuring change with psychometric devices. In *Evaluative research: Strategies and methods.* Washington: American Institute for Research, 1970.

Ebel, R. Criterion-referenced measurements: Limitations. *School Review*, 1971, *79*, 282–288.

Ebel, R. Evaluation and educational objectives. *Journal of Educational Measurement*, 1973, *10*, 273–279.

Ferguson, R. L. A model for computer-assisted criterion-referenced measurement. *Education*, 1970, *91*, 25–31.

Guttman, L. Integration of test design and analysis. In *Proceedings of the 1969 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1969.

Hambleton, R. K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159–170.

Hively, W. Domain-referenced achievement testing. AERA Symposium, 1970, pp. 1–6.

Hively, W.; Patterson, H; & Page, S. A "universe defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, *5*, 275–290.

Macready, G. B., & Merwin, J. C. Homogeneity within item forms in domain referenced testing. *Educational and Psychological Measurement*, 1973, *33*, 351–360.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205–216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education.* Berkeley, Cal.: McCutchan, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurements. In C. W. Harris (Ed.), *Problems in criterion-referenced measurement.* Monograph Series in Evaluation, No. 3. Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, *28*, 95–104.

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

Rosin, A. A computer program to generate exams and homework assignments. *Physics Teacher*, 1974, *12*, 39–41.

Shoemaker, D. M. Criterion-referenced measurement revisited. *Educational Technology*, 1971, *79*, 289–298.

Smythe, M. J.; Kibler, R. J.; & Hutchings, P. W. A comparison of norm-referenced and criterion-referenced measurement with implications for communications instruction. *The Speech Teacher*, 1973, *22*, 1–17.

Swaminathan, H.; Hambleton, R. K.; & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*.

Tinkleman, S. N. Planning the objective test. In R. L. Thorndike (Ed.), *Educational measurement* (2d ed.). Washington: American Council on Education, 1971.