CHAPTER 1
CRITERION-REFERENCED MEASUREMENT: ITS MAIN APPLICATIONS, PROBLEMS AND FINDINGS

Wim J. van der Linden
Technische Hogeschool Twente, Postbus 217, 7500 A E Enschede, The Netherlands

ABSTRACT

The need for criterion-referenced measurements has mainly arisen from the introduction of instructional programs organized according to modern principles from educational technology. Some of these programs are discussed, and it is indicated for what purposes criterion-referenced measurements are used. Three main problems of criterion-referenced measurement are distinguished: The problem of criterion-referenced scoring and score interpretation, the problem of criterion-referenced item and test analysis, and the problem of mastery testing. For each of these problems a variety of solutions and approaches have been suggested. It is the purpose of the paper to provide an overview of these and to introduce the reader to the original literature.

The recent rise of a number of new learning strategies has basically changed the meaning of measurement in education and made new demands on the construction, scoring, and analysis of educational tests. Educational measurements satisfying these demands are usually called criterion-referenced, while traditional measurements are often known as norm-referenced.

The common feature of these learning strategies is their objective-based character. All lead to instructional programs being set up and executed according to well-defined, clear-cut learning objectives. The organizational measures taken to realize these objectives differ, however. For example in learning for mastery, one of the most popular new developments (Block, 1971b; Bloom, 1968; Bloom, Hastings, & Madaus, 1971, chap. 3), the following is done: First, the learning objectives are kept fixed during the implementation of the program. Second, the program is divided into a sequence of small learning units, and students are allowed to proceed to the next unit only after they have mastered the preceding one. Third, end-of-unit tests are administered providing students and instructors with quick feedback. One of the principal uses of these tests is to separate students who master the unit from those who do not (mastery testing). Fourth, when a student does not master the unit, he is given corrective learning materials or remedial teaching. Fifth, as extra learning time is needed for going
through these materials and instruction, mastery learning allows some
differentiation in tempo. There is no differentiation in level, however, since
the learning objectives are kept fixed and equal to all students. (For these
five properties of mastery learning and a further explanation thereof, see
Warries, 1977a.)

The context in which mastery learning seeks to realize its goal is regular
group instruction. In each unit all students receive the same instruction
until the end-of-unit test. Although they have much in common with mastery
learning, individualized instructional systems like the Pittsburgh Individually
Prescribed Instruction (IPI) (Glaser, 1968) and Flanagan's Program for
Learning According to Needs (PLAN) (Flanagan, 1967) differ in this respect.
These systems allow students to reach the learning objectives in different
ways. Each student is provided with his own route through the instructional
units and with learning materials matching his entering behavior and
aptitudes. Individualized instruction has mainly been motivated by the view-
point underlying Aptitude Treatment Interaction (ATI) research (Cronbach &
Snow, 1977), namely that subjects can react differently to treatments and that
treatments which are best on the average may therefore be worst in individual
cases. The methodology needed to test whether ATI's are present has been
reviewed by Cronbach and Snow (1977, chaps. 1-4) and Plomp (1977).
Decision rules for assigning students to treatments are given in van der
Linden (1980d, 1981b) and Vijn (1980).

The most far-reaching mode of individualization takes place in Computer-Aided
Instruction (CAI) (Atkinson, 1968; Suppes, 1966; Suppes, Smith, & Beard,
1975). In CAI the instruction is interactive and each next step depends on
the student's response to the preceding one, thus creating a system in which
the student is, within given boundaries, choosing his own individual
instruction.

Compared with traditional educational approaches, one of the most conspicuous
properties of the testing programs inherent in these modern learning
strategies is their frequency of testing. At several points of time, tests are
involved for several purposes. Beginning-of-unit tests describe the entering
behavior and skills of students who are about to start with the unit. When
the unit itself covers a sequence of objectives, scores on these tests can be
used for deciding where to place students in this sequence. When
individualization is based on ATI research, aptitude tests are usually
administered to assign students to those treatments that promise best results.

Once the student has taken up the unit, intermediate tests can be
administered for monitoring his learning. The principal use of these tests is
to describe relevant aspects of the learning process enabling students and
instructor to re-adjust it when necessary. This intermediate use of tests is
known as formative evaluation.

Tests are also used as end-of-unit tests for describing the level of master the
student has attained when he has completed the unit. If the information from
these tests serves as a basis for deciding whether the student has mastered
the unit and can be moved up to the next unit, the testing procedure is
known as mastery testing. End-of-unit tests may be used for diagnostic purposes as well; then they indicate the objectives for which the student's performances are poor, and which part of the unit he has to cover again.

In all these modes of testing, scores are used only for instructional purposes. In particular, they are not used for grading students (summative evaluation). For that purpose a separate test is ordinarily administered at the end of the program, covering the content of the units but ignoring the unit test scores given earlier.


As already indicated, the introduction of such strategies as learning for mastery and individualized instruction has led to a change in the use and the interpretation of test scores. The modes of testing outlined above are all concerned with a behavioral description of students. By doing so, it is possible to control the learning process and to make optimal decisions, for example, on the placement of students in instructional units and their end-of-unit mastery level. Traditional testing procedures are, however, better suited for differentiating between subjects, and mostly serve as an instrument for (fixed-quota) selection. The psychometric analysis of these tests is generally adapted to this use.

The attempts to remove traditional assumptions from educational testing and to replace these by assumptions better adjusted to the use of tests in modern learning strategies arose in the beginning of the sixties. As a result the term "criterion-referenced measurement" has been coined. Elsewhere (van der Linden, 1979) we have given a review in which three main problems of criterion-referenced measurement are distinguished, each for the first time signaling a different "historical" paper. These problems are: The problem of criterion-referenced scoring and score interpretation, the problem of criterion-referenced item and test analysis, and the problem of mastery decisions. Following are short introductions to these three problems.

**CRITERION-REFERENCED SCORING AND SCORE INTERPRETATION**

Glaser (1963), in his paper on instructional technology and the measurement of learning outcomes, confronted two possible uses of educational tests and their areas of application. The first is that tests can supply norm-referenced measurements. In norm-referenced measurement the performances of subjects are scored and interpreted with respect to each other. As the name indicates, there is always a norm group, and the interest is in the relative standing of the subjects to be tested in this group. This finds expression in scoring methods as percentile scores, normalized scores, and age equivalents. Tests are constructed such that the relative positions of subjects come out as reliably as possible. An outstanding example of an area where norm-
referenced measurements are needed is testing for selection (e.g., of applicants for a job). In such applications the test must be maximally differentiating in order to enable the employer to select the best applicants.

The second use is that tests can supply criterion-referenced measurements. In criterion-referenced measurement the interest is not in using test scores for ranking subjects on the continuum measured by the test, but in carefully specifying the behavioral referents (the "criterion") pertaining to scores or points along this continuum. Measurements are norm-referenced when they indicate how much better or worse the performances of individual subjects are compared with those of other subjects in the norm group; they are criterion-referenced when they indicate what performances a subject with a given score is able to do, and what his behavioral repertory is, without any reference to scores of other subjects. This descriptive use of test scores is needed in the testing programs of the learning strategies mentioned earlier. For a further clarification of the distinction between norm-referenced and criterion-referenced test score interpretations, we refer to Block (1971a), Carver (1974), Ebel (1962, 1971), Flanagan (1950), Glaser and Klaus (1962), Glaser and Nitko (1971), Glass (1978), Hambleton, Swaminathan, Algina, and Coulson (1978), Linn (1980), Nitko (1980), and Popham (1978).

How to establish a relation between test scores and behavioral referents is what can be called the problem of criterion-referenced score interpretation. We have also called this the problem of local validity, since Glaser (1963) goes beyond the classical validity problem and does not ask for the validity of a test, which is the classical question, but for the validity of the interpretation of test scores. The validity of the test must, as it were, be locally specifiable for points on the continuum underlying the test (van der Linden, 1979).

Several answers have been proposed to the problem of criterion-referenced score interpretation, some of which will be mentioned here. One is a constructive approach based on the idea of referencing test scores to a domain of tasks (e.g., Hively, 1974; Hively, Maxwell, Rabehl, Sension, & Lundin, 1973; Hively, Patterson, & Page, 1968; Osburn, 1968). In this approach, the test is a random sample from a domain of tasks (or may be conceived of as such) defined by a clear-cut learning objective and test scores are interpreted with respect to this domain. Domain-referenced testing usually involves the use of binomial test models and has been the most popular way of criterion-referencing so far. Another approach is an empirical method in which the congruence between items and objectives is determined by assessing how sensitive the items are to objective-based instruction. The objectives are then used to interpret the test performances. An example of this approach is Cox and Vargas' (1966) pretest-posttest method, which has led to a multitude of variants and modifications (for a review, see Berk, 1980b; van der Linden, 1981a). A third approach is subjective and uses content-matter specialists to judge the congruence between items and learning objectives (Hambleton, 1980; Rovinelli & Hambleton, 1977). Yet another approach has been followed by Cox and Graham (1966) who conceived the continuum to be referenced as a Guttman scale and used scalogram analysis to scale items pertaining to a sequence of learning objectives. By so doing,
they were able to predict for points along the scale which arithmetical skills the students in their example were able to perform. Wright and Stone (1979, chap. 5) used the Rasch model to link points on the continuum underlying the test to behaviors. The crucial difference between scalogram analysis and this approach is that the former assumes Guttman item characteristic curves and allows only deterministic statements about behaviors whereas use of the Rasch model entails probabilistic interpretations.

The above five approaches were mentioned because they have been most popular so far or indicate promising developments. For a more exhaustive review, we refer to Nitko (1980).

CRITERION-REFERENCED ITEM AND TEST ANALYSIS

Popham and Husek's (1969) paper is a second milestone in the history of criterion-referenced measurement. It goes beyond Glaser's paper in that it adds a distinction in item and test analysis to norm-referenced and criterion-referenced measurement. According to Popham and Husek, both types of measurement differ so basically that classical models and procedures for item and test analysis are inadequate for criterion-referenced measurement and their outcomes sometimes even misleading.

The key-word in the analysis of items and tests for norm-referenced measurement, with its emphasis on differentiating between subjects, is variance. Classical models and procedures rely on the presence of a large amount of variability of scores. In criterion-referenced measurement, however, this condition will seldom be fulfilled because it is not the variability of scores which is critical but their relation to a criterion. In this connection, Popham and Husek refer to classical reliability analysis. Consistency, both internally and temporarily, is a desirable property not only of norm-referenced but of criterion-referenced measurements as well. Nevertheless, classical reliability coefficients are variance-dependent and thus often low for criterion-referenced measurements. For this reason, they plead for new models and procedures for item and test analysis. These must be models and procedures which, more than classical test theory, make allowance for the specific requirements criterion-referenced measurement imposes on item construction, test composition, scoring, and score interpretation.

Popham and Husek's plea has led to a variety of proposals. A proposal to adapt classical test theory for use with criterion-referenced measurement is due to Livingston (1972a). He argued that in criterion-referenced measurement we are no longer interested in estimating the deviations of individual true scores from the mean but in estimates of deviations from the cut-off score. Central in his approach is a criterion-referenced reliability coefficient defined as the ratio of true to observed score variance about the cut-off score. Livingston's proposal stimulated others to contribute (Harris, 1972; Lovett, 1978; Shavelson, Block, & Ravitch, 1972; see also Livingston, 1972b, 1972c, 1973), and has had the merit of making a larger public aware of the issue of criterion-referenced test analysis.
As noted above, random sampling of tests from item domains has been the most popular way of criterion-referencing test scores. However, when tests are sampled and the concern is with domain scores and not with true test scores, classical test theory does not hold (unless all items in the domain have equal difficulty). Brennan and Kane (1977a) proposed to use generalizibility theory when domain sampling takes place and transformed Livingston's reliability coefficient into what they called an index of dependability. At the same time they presented a modification of this index which can be interpreted as a signal-noise ratio for domain-referenced measurements, while later two general agreement coefficients were given from which these indices can be derived as special instances (Brennan & Kane, 1977b; Kane & Brennan, 1980). A summary of these developments is given in Brennan (1980).

In the foregoing approaches the emphasis is on developing test theory reflecting the reliability or dependability with which deviations of individual true or domain scores from the cut-off score can be estimated. It can be argued, however, that when criterion-referenced tests are used for mastery decisions the concern should not be so much with these deviations as with the "reliability" and "validity" of the decisions. Since approaches along this line belong to the content of the next section, we will postpone a discussion of their results until then.

The first proposal of a criterion-referenced item analysis was made by Cox and Vargas (1966). As already indicated in the preceding section, their pretest-posttest method of item validation is based on the idea that criterion-referenced items ought to be sensitive to objective-based instruction. It measures this sensitivity by the difference in item p-value before and after instruction. The rationale of the method is discussed in Coulson and Hambleton (1974), Cox (1971), Edmonston and Randall (1972), Hambleton and Gorth (1971), Henrysson and Wedman (1973), Millman (1974), Roudabush (1973), Rovinelli and Hambleton (1977), and Wedman (1973, 1974a, 1974b). Slightly different approaches can be found in Brennan and Stolurow (1971), Harris (1976), Herbig (1975, 1976), Kosecoff and Klein (1974), and Roudabush (1973). Popham (1971) offers a chi-squared test for detecting items with atypical differences in p-value. Sauer's (1966) correlation between item and total change score is often considered the counterpart of the norm-referenced item-test correlation.

A critical review of item analysis based on the pretest-posttest method is given in van der Linden (1981a) who also offers an alternative derived from latent trait theory. (See also van Naerssen, 1977a, 1977b.) Other reviews of criterion-referenced item analysis procedures are given in Berk (1980b) and Hambleton (1980). For reviews of test analysis procedures, we refer to Berk (1980a), Hambleton, Swaminathan, Algina, and Coulson (1978), and Linn (1979).

Popham and Husek's diagnosis of the role of test score variance in the analysis of criterion-referenced measurement has been a hotly-debated issue (Haladyna, 1974a, 1974b; Millman & Popham, 1974; Simon, 1969; Woodson, 1974a, 1974b). Our own opinion is that, although it has given rise to many
worthwhile developments and comes close to a fundamental problem in test theory, it is erroneous as far as the classical test model is concerned. Apart from the requirement of finite variance, this model does not contain any assumption with respect to score variance. (A full statement of these assumptions can be found, e.g., Lord & Novick, 1968, section 3.1.) Low variability of test scores in criterion-referenced measurement can therefore never invalidate the classical test model. Latent trait theory has called attention to the more fundamental problem of population-dependent item and test parameters and offers models in which these are replaced by parameters being not only variance-independent but independent of any distributional characteristic (e.g., Birnbaum, 1968; Lord, 1980; Wright & Stone, 1979; for an informal introduction, see van der Linden, 1978b). In fact, Popham and Husek's paper alludes to this problem but wrongly claims it as an exclusive problem of criterion-referenced measurement.

MASTERY DECISIONS

Hambleton and Novick (1973) were the first to introduce a decision-theoretic approach to criterion-referenced measurement. In modern learning strategies, such as the ones briefly reviewed above, the ultimate use of criterion-referenced measurements is mostly not measuring students but making instructional decisions. Hambleton and Novick compare the use of norm-referenced and criterion-referenced measurements with the concepts of fixed-quota and quota-free selection (Cronbach & Gleser, 1965).

Norm-referenced tests are needed to differentiate between subjects in order to select a predetermined number of best performing persons, regardless of their actual level of performance (fixed-quota selection). Criterion-referenced tests are mostly used to select all persons exceeding a performance level fixed in advance, regardless of their actual number (quota-free selection). With criterion-referenced tests, this fixed level of performance is known as the mastery score, and selection with this score as mastery testing.

It is important to note that in mastery testing there is thus only one point on the criterion-referenced continuum that counts: the mastery score. This score divides the continuum in mastery and non-mastery areas.

There is an alternative conception of mastery which is not based on a continuum model, as the former, but on a state model. In this conception mastery and non-mastery are viewed as two latent states, each characterized by a different set of probabilities of a successful response to the test items. There is no state between mastery and non-mastery and it is not necessary to set a mastery score, as with continuum models. By fitting the model to test data it is left to nature to define who is a master and who is not. Relevant references to the literature on state model are Bergan, Cancelli, and Luiten (1980), Besel (1973), Dayton and Macready (1976, 1980), Emrick (1971), Emrick and Adams (1969), Macready and Dayton (1977, 1980a, 1980b), Reulecke (1977a, 1977b), van der Linden (1980c, 1981c, 1981d, 1981e), Wilcox
(1979a), and Wilcox and Harris (1977). For a discussion of the differences between continuum and state models, we refer to Meskauskas (1976) and van der Linden (1978a).

Though mastery is defined using true scores or states, mastery decisions are made with test scores containing measurement errors. Due to this, decision errors can be made, and a student can be wrongly classified as a master (false positive errors) or a non-master (false negative errors). An important mastery testing problem is how to choose a cut-off score on the test so that decisions are made as optimally as possible. A second problem is the psychometric analysis of mastery decisions. Classical psychometric analyses view tests as instruments for making measurements, and this point of view has pervaded its models and procedures. When tests are used for making decisions, this is no longer correct, however.

Hambleton and Novick (1973) proposed the use of Bayesian decision theory to optimize the cut-off score on the test. Important in applying decision theory is the choice of the loss function representing the "seriousness" of the decision outcomes on a numerical scale. Several have been used so far: threshold loss, linear loss, and normal ogive loss functions (for a comparison, see van der Linden, 1980a). The threshold loss function is used, e.g., in Hambleton and Novick (1973), Huynh (1976), and Mellenbergh, Koppelaar, and van der Linden (1977). Livingston (1975) and van der Linden and Mellenbergh (1977) show how semi-linear, respectively, linear loss functions can be used to select optimal cut-off scores. The use of normal ogive loss function is suggested by Novick and Lindley (1978).

All these applications of decision theory to the optimal cut-off score problem can be used with a Bayesian as well as an empirical Bayes interpretation. Fully Bayesian approaches are presented and discussed in Hambleton, Hutten, and Swaminathan (1976), Lewis, Wang, and Novick (1975), and Swaminathan, Hambleton, and Algina (1975).

It is also possible to base the making of mastery decisions on the Neyman-Pearson theory of testing statistical hypotheses. Then no explicit loss functions are chosen but the consequences of false positive and false negative errors are evaluated via determining the sizes of the type I and type II errors of testing the hypothesis of mastery. Approaches along this line, all based on binomial error models, can be found in Fhaner (1974), Klauer (1972), Kriewall (1972), Millman (1973), van den Brink and Koele (1980), and Wilcox (1976).

A minimax solution to the optimal cut-off score problem, which resembles the (empirical) Bayes approach in that the losses must be specified explicitly but does not assume the availability of (subjective) information about true score, is presented in Huynh (1980a).

The problem of the analysis of mastery decisions was also addressed by Hambleton and Novick (1973). They proposed to determine the reliability of mastery decisions by assessing the consistency of decisions from test-retest or parallel test administrations using the coefficient of agreement. Similarly,
an independently measured criterion could be used for the determination of the validity of mastery decisions. Swaminathan, Hambleton, and Algina (1974) suggested to replace this coefficient, which was already proposed for this purpose by Carver (1970), by the chance-corrected coefficient kappa of Cohen (1960). These two coefficients suppose the availability of two test administrations. Both Huynh (1976) and Subkoviak (1976) presented a single administration method of estimating the coefficients, which were derived using the beta-binomial model and an equivalent set of assumptions, respectively. Another single administration method was given by Marshall and Haertel (1976). All these methods have been extensively examined and compared to each other: Algina and Noe (1978), Berk (1980a), Divgi (1980), Huynh (1976, 1979), Huynh and Saunders (1980), Marshall and Serlin (1979), Peng and Subkoviak (1980), Subkoviak (1978, 1980), Traub and Rowley (1980), and Wilcox (1979c).

The idea that the "reliability" of decisions can be determined via their consistency goes back to the assumption that classical test theory holds for decisions just as it does for measurements. Mellenbergh and van der Linden (1979) have shown that the assumption is incorrect and that test-retest or parallel test consistency of decisions does not necessarily reflect their accuracy. They recommend the use of coefficient delta (van der Linden & Mellenbergh, 1978) which indicates how optimal the decisions actually made are with respect to the true mastery and non-mastery states. Comparable approaches have been undertaken by de Gruijter (1978), Livingston and Wingersky (1979), and Wilcox (1978d).

Also relevant to the issue of mastery setting with continuum models are techniques for setting mastery scores. These techniques all somehow translate the learning objectives into a mastery score on the true score continuum underlying the test. Hence, they should be distinguished from the decision-theoretic approaches mentioned above which, once the mastery score on the continuum has been set, indicate how to optimally select the cut-off score on the test. Several techniques of setting mastery scores have been proposed. Reviews of these techniques are given, e.g., in Glass (1978), Hambleton (1980), Jaeger (1979), and Shepard (1979, 1980). An approach accounting for possible uncertainty in setting standards is given in de Gruijter (1980).

CONCLUDING REMARKS

In this paper we have presented an overview of the way criterion-referenced measurement is used in modern instructional programs, what its main problem areas are, and how these have been approached. The purpose of the paper was to summarize developments and results and to provide an introduction to the original literature.

It should be noted, however, that not all aspects of criterion-referenced measurement have been considered. For example, we did not refer to

This paper has thus not touched on all aspects of criterion-referenced measurement. Yet, we hope it has given an impression of the directions the field of criterion-referenced measurement is taking. Three "historical" papers have guided us in exploring the field. We were able to observe that a great variety of solutions have been formulated in response to these papers. Most of the solutions promise important improvements of educational testing technology and, thereby, of educational practice.

REFERENCES


Harris, N.D.C. An index of effectiveness for criterion-referenced items used in pre-tests and post-tests. *Programmed Learning and Educational Technology, 11,* 125-132, 1974.


Livingston, S.A. Reply to Shavelson, Block, and Ravitch's "Criterion-referenced testing: Comments on reliability". Journal of Educational Measurement, 9, 139-140, 1972c.


van der Linden, W.J. On the estimation of the proportion of masters in criterion-referenced measurement. Submitted for publication, 1981d.

van der Linden, W.J. The use of moment estimators for mixtures of two binomials with one known success parameter. Submitted for publication, 1981e.


