ACT RESEARCH REPORT

No. 53

53

September 1972

TOWARD AN INTEGRATION OF THEORY AND METHOD FOR CRITERION-REFERENCED TESTS

R. K. Hambleton
M. R. Novick

PUBLISHED BY THE RESEARCH AND DEVELOPMENT DIVISION

THE AMERICAN COLLEGE TESTING PROGRAM

P. O. BOX 168, IOWA CITY, IOWA 52240

TOWARD AN INTEGRATION OF THEORY AND METHOD FOR CRITERION-REFERENCED TESTS

Prepared by the Research and Development Division
The American College Testing Program

For additional copies write:
Publication and Information Services Division
The American College Testing Program
P.O. Box 168, Iowa City, Iowa 52240

(Check or money order must accompany request.) Price: $1.00

ABSTRACT

In this paper, an attempt has been made to synthesize some of the current thinking in the area of criterion-referenced testing as well as to provide the beginning of an integration of theory and method for such testing. Since criterion-referenced testing is viewed from a decision-theoretic point of view, approaches to reliability and validity estimation consistent with this philosophy are suggested. Also, to improve the decision-making accuracy of criterion-referenced tests, a Bayesian procedure for estimating true mastery scores has been proposed. This Bayesian procedure uses information about other members of a student's group (collateral information), but the resulting estimation is still criterion-referenced rather than norm-referenced in that the student is compared to a standard rather than to other students. In theory, the Bayesian procedure increases the "effective length" of the test by improving the reliability, the validity, and more importantly, the decision-making accuracy of the criterion-referenced test scores.

TOWARD AN INTEGRATION OF THEORY AND METHOD FOR CRITERION-REFERENCED TESTS[1,2]

Ronald K. Hambleton
Melvin R. Novick

Over the years, standard procedures for constructing, administering, and analyzing tests, and interpreting scores in the context of standard instructional models and methods have become well-known to educators. With these models, tests have been used primarily and most successfully to estimate each examinee's ability level and to permit comparative statements (e.g., ranking) across examinees. Recently, however, there have been numerous suggestions for, and demonstrations of, instructional models and methods in the schools where the well-known classical mental test models for test construction and test score interpretation appear to be less useful. Examples of these instructional models include: *Computer-Assisted Instruction* (Atkinson, 1968; Suppes, 1966), *Individually Prescribed Instruction* (Glaser, 1968), *Project PLAN* (Flanagan, 1967, 1969), and *A Model of School Learning* (Carroll, 1963, 1970; Bloom, 1968; Block, 1971). Common to most of these instructional models as well as to several others are such features as the specification of the curriculum in terms of behavioral objectives, detailed diagnosis of beginning students, the availability of multiple instructional modes, individual pacing and sequencing of material, and the careful monitoring of student progress.

While not all educators agree on the usefulness of these instructional models in the schools, the position taken in this paper is that these models are useful, and that their usefulness will be enhanced by developing testing methods and decision procedures specifically designed for use within the context of these models. The purpose of this paper is to outline some appropriate statistical methods that may prove of use in making instructional decisions for students.

It appears that much of the discussion in this area (for example, see Block, 1971; Carver, 1970; Ebel, 1971; and Glaser & Nitko, 1971) stems from different understandings as to the basic purpose of testing in these instructional models. It would seem to us that in most cases the pertinent question is whether or not the individual examinee has attained some prescribed degree of competence on an instructional performance task (see, for example, Harris, 1972b). Questions of precise achievement levels and comparisons among individuals on these levels seem to be largely irrelevant. In many of the new instructional models, tests are used to determine on which instructional objectives an examinee has met the acceptable performance level standard set by the model designer. This test information is usually used immediately to evaluate the student's mastery of the instructional objectives covered in the test, so as to appropriately locate him for his next instruction (Glaser & Nitko, 1971). Tests especially designed for this particular purpose have come to be known as *criterion-referenced tests*. Criterion-referenced tests are specifically designed to meet the measurement needs of the new instructional models. In contrast, the better known *norm-referenced tests* are principally designed to produce test scores suitable for ranking individuals on the ability measured by the test. Sometimes this occurs with the understanding that some cut-off score will be introduced to reject some percentage of students for the next level of instruction.

---

[1]This paper was begun while the first author was on a Summer Postdoctoral Fellowship at The American College Testing Program and completed with support from the Office of Education. The research reported herein was performed pursuant to Grant No. OEG-0-72-0711 from the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy. The authors are indebted to Roy Williams and Thomas Hutchinson for helpful comments.

[2]Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1972.

Criterion-Referenced Tests: Definitions and Selected Issues

A "criterion-referenced test" has been defined in a multitude of ways in the literature. (See, for example, Glaser & Nitko, 1971; Harris & Stewart, 1971; Ivens, 1970; Kriewall, 1969; and Livingston, 1972a). The definitions are sufficiently different that a test may be classified as norm-referenced according to one definition, criterion-referenced according to another, or more typically, exhibit characteristics of each to a greater or lesser extent depending on the definition. The intentionally most restrictive definition of a criterion-referenced test was proposed by Harris and Stewart (1971): "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performance, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed." On the other hand, possibly the least restrictive definition is that by Ivens (1970) who defined a criterion-referenced test as "one made up of items keyed to a set of behavioral objectives." A very flexible definition has been proposed by Glaser and Nitko (1971): "A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards." According to Glaser and Nitko, "The performance standards are usually specified by defining some domain of tasks that the student should perform. Representative samples of tasks from this domain are organized into a test. Measurements are taken and are used to make a statement about the performance of each individual relative to that domain." This definition is less restrictive than Harris and Stewart's in that it does not limit consideration to a single instructional objective. A common thread running through the various approaches to criterion-referenced tests is that the definition of a well-specified content domain and the development of procedures for generating appropriate samples of test items are important. (For more on this, see, Bormuth, 1970; Glaser & Nitko, 1971; and Hively, Patterson, & Page, 1968.)

It should be noted that these are also concerns of those interested in constructing norm-referenced tests; however, not to the same extent. Less often is there an interest in making inferences about which *particular* skills an individual has or does not have from his performance on a norm-referenced test. Thus, norm-referenced testing is seldom diagnostic. Primary examples would be the Scholastic Aptitude Test (SAT) and, to a lesser extent, the ACT Assessment. Exceptions would be tests such as the *Iowa Tests of Basic Skills* which have important features of both norm- and criterion-referenced tests. Such tests are norm-referenced because they are geared to reporting how well a student compared with others in certain well-defined populations (e.g., through percentile scores). Yet, they are criterion-referenced in that they are keyed to specific instructional objectives, are multiscaled, and diagnostic. However, they do not involve apriori judgment as to acceptable performance levels and a consequent judgment as to whether or not an individual student attains this performance level. Further distinctions between norm-referenced tests and criterion-referenced tests have been presented by Block (1971), Ebel (1971), Glaser (1963), Glaser and Nitko (1971), Hambleton and Gorth (1971), Hieronymus (1972), and Popham and Husek (1969).

If one accepts the Glaser and Nitko definition of a criterion-referenced test, it is apparent that the test may often be multidimensional while made up of unidimensional subscales. That is, the items from a criterion-referenced test are organized in distinct and different subscales of homogeneous items measuring common skills. (The possibility of a single item subscale is not ruled out.) An instructional decision for each individual is then often made on the basis of his performance on *each* subscale. Major interest may, thus, rest on the reliability and validity of subscale scores.

One of the problems yet to be reckoned with for criterion-referenced tests is an instance of the bandwidth-fidelity issue (Cronbach & Gleser, 1965). When the total testing time is fixed and there is interest in measuring many competencies, one may be faced with the problem of whether to obtain very precise information about a small number of competencies or less precise information about many more competencies. Time allocation algorithms (analytical procedures for deciding

how many items on a test should measure each objective) of a rather different kind than those presented by Woodbury and Novick (1968) and Jackson and Novick (1970) will be required. They will be closer in spirit, but not identical to those given by Cronbach and Gleser (1965). The problem of how to fix the length of each subscale so as to maximize the percentage of correct decisions or some similar measure of overall decision-making accuracy on the basis of test results has yet to be resolved or, indeed, to be formulated satisfactorily.

Distinction among Testing Instruments, Measurement, and Decisions

Some clarification concerning appropriate measurement models for these new instructional programs can be obtained by properly distinguishing between testing instruments and measurement. With the availability of a test theory for norm-referenced measurement (e.g., see Lord & Novick, 1968), we have procedures for constructing appropriate measuring instruments, i.e., norm-referenced tests. Then, the pertinent question seems to be whether or not the instructional models which require different kinds of measurements (i.e., criterion-referenced measurement) also require new kinds of tests or whether the usual tests with alternate procedures for interpreting test scores can be used. We subscribe to the belief that different tests are needed, constructed to meet quite different specifications than those typically set for norm-referenced tests (Glaser, 1963). We do not propose, however, to explicate a developed theory of criterion-referenced measurement in this paper nor to prescribe a technology for criterion-referenced test development. Such explication should be based both on a well-developed instructional theory and on a decision-theoretic formulation of the measurement problem. Only the latter is even touched on here. The test development technology would be concerned primarily with methods of obtaining a representative sample of behaviors from a specified domain.

It should be noted that a norm-referenced test can be used for criterion-referenced measurement, albeit with some difficulty, since the selection of items is such that many objectives will very likely not be covered on the test or, at best, will be covered with only a few items. A criterion-referenced test constructed by procedures especially designed to facilitate criterion-referenced measurement can and sometimes is used to make norm-referenced measurements. However, a criterion-referenced test is not constructed specifically to maximize the variability of test scores (whereas a norm-referenced test is). Thus, since the distribution of scores on a criterion-referenced test will tend to be homogeneous, it is obvious that such a test will be less useful for ordering individuals on the measured ability. In summary, then, a norm-referenced test can be used to make criterion-referenced measurements, and a criterion-referenced test can be used to make norm-referenced measurements, but neither usage will be particularly satisfactory.

Thus, it may be misleading to talk about tests as either norm-referenced or criterion-referenced since measurements obtained from either testing instrument can be explained with a norm-referenced interpretation, criterion-referenced interpretation, *or* both. The important distinction, we believe, is between norm-referenced measurement and criterion-referenced measurement. This distinction was made by Glaser (1963) but seems to have been ignored by several subsequent writers.

Decision-Theoretic Approach to Criterion-Referenced Measurement

Our own conceptual framework for criterion-referenced measurement goes this way. Like Cronbach and Gleser (1965), we see testing as a decision-theoretic process. One of the main differences between norm-referenced tests and criterion-referenced tests is in terms of the kinds of

decisions they are specifically designed to make. Norm-referenced measurement is particularly useful in situations where one is interested in "fixed-quota" selection or ranking of individuals on some ability continuum. Criterion-referenced measurement involves what Cronbach and Gleser (1965) would call a "quota-free" selection problem. That is, there is no quota on the number of individuals who can exceed the *cut-off scores* or *threshold* on a criterion-referenced test. A cut-off score is set for each subscale of a criterion-referenced test to separate examinees into two mutually exclusive groups. One group is made up of examinees with high enough test scores ($\geq$ the cut-off score) to infer they have mastered the material to a desired level of proficiency. The second group is made up of examinees who did not achieve the minimum proficiency standard. At this stage of the development of a theory of criterion-referenced measurement, the establishment of cut-off scores is primarily a value judgment. Much research might usefully be undertaken to provide guidelines for this judgment. The educational goal is, of course, to have everyone achieve the standards. This is attempted by means such as individualizing instruction to the point of providing multiple instructional modes (Cronbach, 1967), individual pacing and sequencing, as well as providing various remedial programs.

The primary problem in the new instructional models, such as individually prescribed instruction, is one of determining, if π_i, the student's mastery level, is greater than a specified standard π_0. Here, π_i is the "true" score for an individual i in some particularly well-specified content domain. It *may* represent the proportion of items in the domain he could answer successfully. Since we cannot administer all items in the domain, we sample some small number to obtain an estimate of π_i, represented as $\hat{\pi}_i$. The value of π_0 is the somewhat arbitrary threshold score used to divide individuals into the two categories described earlier, i.e., Masters and Nonmasters.

Basically then, the examiner's problem is to locate each examinee in the correct category. There are two kinds of errors that occur in this classification problem: false positives and false negatives. A false-positive error occurs when the examiner estimates an examinee's ability to be above the cutting score when, in fact, it is not. A false-negative error occurs when the examiner estimates an examinee's ability to be below the cutting score when the reverse is true. The seriousness of making a false-positive error depends to some extent on the structure of the instructional objectives. It would seem that this kind of error has the most serious effect on program efficiency when the instructional objectives are hierarchical in nature. On the other hand, the seriousness of making a false-negative error would seem to depend on the length of time a student would be assigned to a remedial program because of his low test performance. (Other factors would be the cost of materials, teacher time, facilities, etc.) The minimization of expected loss would then depend, in the usual way, on the specified losses and the probabilities of incorrect classification. This is then a straightforward exercise in the minimization of what we would call *threshold* loss.

In an attempt to view the above discussion in a more formal manner, suppose we take some cutting score, π_0, and define a parameter ω such that

$$\omega = 1 \text{ if } \pi \geq \pi_0$$

$$\omega = 0 \text{ if } \pi < \pi_0 .$$

Now if we obtain an estimate of π_i, then an estimate of ω can be obtained in the following way:

$$\hat{\omega} = 1, \text{ if } \hat{\pi}_i \geq \pi_0 \text{ and}$$

$$\hat{\omega} = 0, \text{ if } \hat{\pi}_i < \pi_0 .$$

Defining our error of estimation as $(\hat{\omega} - \omega)$, it is clear that the error takes on one of three values, $+1$, -1, 0, corresponding to whether we make a false-positive error, a false-negative error, or a correct classification. Also, note that the squares of the errors and their absolute values are identical. Thus, any procedure that minimizes squared-error loss (SEL) in the ω metric also minimizes absolute-error loss (AEL) in that metric. Furthermore, the minimization of SEL and AEL in the ω metric is equivalent to the minimization of threshold loss for π in the special case where the losses associated

with false positives and false negatives are equal. The criterion-referenced measurement problem is, thus, one of determining an estimator $\hat{\omega}$ of ω by determining an estimator $\hat{\pi}$ of π with a *threshold loss* function and converting this to an estimate of ω. We shall exemplify this process shortly. Note that with threshold loss, the estimate $\hat{\pi}$ of π is not a single number but one of two intervals, $[0, \pi_0)$ or $[\pi_0, 1]$. It might well be argued that what we describe here is not "measurement" at all; and, in fact, it might be useful to avoid use of the term measurement in the above context.

The following example will illustrate an application of threshold loss. To estimate a person's π value under threshold loss, first write down the losses associated with the two kinds of incorrect decisions. Thus, we take

$$\ell(e) = 0 \qquad \text{if } e = 0,$$

$$\ell(e) = a > 0 \qquad \text{if } e = +1,$$

$$\ell(e) = b > 0 \qquad \text{if } e = -1 .$$

The expected loss if we set $\hat{\omega} = 1$ is

$$a[\text{Prob}(\pi < \pi_0|\text{data})], \qquad (1)$$

and if we set $\hat{\omega} = 0$, it is

$$b[\text{Prob}(\pi \geq \pi_0|\text{data})]. \qquad (2)$$

Thus, we set $\hat{\omega} = 1$ or 0 depending upon whether expression (1) or expression (2) is the smaller. This decision corresponds to estimating with threshold loss whether $\pi \geq \pi_0$ or $\pi < \pi_0$. Note, however, that we may decide that $\omega = 0 (\pi < \pi_0)$, i.e., take $\hat{\omega} = 0$ not because $\text{Prob}(\pi < \pi_0|\text{data}) > \text{Prob}(\pi \geq \pi_0|\text{data})$ but because a is very much larger than b, the loss associated with a false positive is very much greater than that associated with a false negative.

Suppose we judge the loss associated with a false positive to be a = 8 "units" and the loss associated with a false negative to be b = 1 unit. Further, suppose that given the data

$$\text{Prob}(\pi \geq \pi_0) = .85 \text{ and, hence, } \text{Prob}(\pi < \pi_0) = .15$$

then, the value of (1) is

$$a[\text{Prob}(\pi < \pi_0)|\text{data}] = (8)\ (.15) = 1.2,$$

and the value of (2) is

$$b[\text{Prob}(\pi \geq \pi_0)|\text{data}] = (1)\ (.85) = .85 .$$

Hence, we take $\hat{\omega} = 0$ and classify the student as a nonmaster. Now, notice that the comparison of (1) and (2) is equivalent to the comparison of the a/b to the ratio

$$[\text{Prob}(\pi \geq \pi_0|\text{data}] \ / \ [1 - \text{Prob}(\pi \geq \pi_0|\text{data})] .$$

This spotlights the fact that the educator need not stipulate a and b in any absolute value. He need only stipulate the ratio a/b. In this example, since $\text{Prob}(\pi \geq \pi_0) = .85$, the student will be classified as a nonmaster unless the ratio $a/b \leq 5.67$. Generally with a and b as given, a student will be classified as a master only if $\text{Prob}(\pi \geq \pi_0) > .89$, approximately.

It should be noted that the above approach generalizes quite easily to situations where there are possibly several different treatments, several relevant levels of mastery on each skill, and several different prerequisite skills. Details of such situations will be given elsewhere.

Bayesian Estimation of Mastery Scores

In order to determine if an examinee has mastered a particular skill (i.e., instructional objective), we analyze his responses to items on a criterion-referenced test designed to measure that skill. These items plus the items designed to measure achievement of other skills are organized together to form a criterion-referenced test.

Each student is assumed to have some mastery score, π_i, which may be the proportion of items in the domain he can answer correctly. The measurement problem is to estimate π_i from some usually small number of test items. Typically, a student's mastery score is estimated to be his proportion-correct score. Mastery scores are estimated for the

purpose of decision-making: If $\hat{\pi}_i \geq \pi_0$, the student is sent on to new work; otherwise with $\hat{\pi}_i < \pi_0$, he is assigned some remedial work. Before presenting a Bayesian solution to the mastery assessment problem, let us consider the problem of estimating a single student's true score π.

Generally, the method of using the proportion-correct as an estimate of π_i is not entirely satisfactory when the number of items on which the proportion is based is few and when there are many students. In situations where one is interested in estimating many parameters; some, by chance, will be substantially overestimated and others, underestimated. The implication of this is that many errors of classification will be made. In estimation or in making mastery decisions on the basis of small amounts of information, we run the risk of making many errors. What is the solution? Because of the extensive amount of testing taking place, it is usually impractical to consider lengthening the test. However, a Bayesian estimation procedure proposed by Novick, Lewis, and Jackson (1972) provides, at least theoretically, a way of obtaining more information on each examinee without requiring the administration of any additional test items. According to Novick et al. (1972), this can be done by using not only the *direct information* provided by a student's (subscale) score, but also the *collateral information* contained in the test data of other students. (Another possibility and worthy of further research is the possibility of using the student's other subscale scores and previous history as collateral information.)

A familiar example of how this can be done comes from the application of classical test theory (Lord & Novick, 1968) to norm-referenced measurement. Within the classical test theory model, each examinee's observed score x on a test may be used as an estimate of his true score τ. The standard deviation of error scores across examinees in the population (standard error of measurement) will be $\sigma_x(1 - \rho_{xx'})^{1/2}$ where σ_x is the standard deviation of observed scores, and $\rho_{xx'}$ is the reliability of the test. This formula provides a measure of the inaccuracy, on the average, of using the observed score as an estimate of true score. An alternative method of estimating true score is to use a regression estimate $\hat{\tau} = x\rho_{xx'} + \mu_x(1 - \rho_{xx'})$, where μ_x is the mean-observed score in the population of examinees. It can be shown that the average error in the population obtained by using $\hat{\tau}$ as an estimator of τ is $\sigma_x\rho_{xx'}^{1/2}(1 - \rho_{xx'})^{1/2}$. This is called the standard error of estimation. By comparing formulas, it is easily seen that the standard error of estimation is smaller than the standard error of measurement and is substantially smaller than the latter when $\rho_{xx'}$ is low. This is because, in effect, we are using information about the group of which the individual is a member to provide "prior" information for the Bayesian estimation of each person's true mastery score. With this approach, under common circumstances, the Bayesian method can effect an increase of precision equivalent to that which would be obtained by adding between 6 and 12 items to the test (see Novick, Lewis, & Jackson, 1972). Thus, the Bayesian method has something substantial to offer in the context of norm-referenced measurement problems, and similarly, it would seem that the same potential exists with criterion-referenced testing problems.

However, it should be noted that our previous discussion has stressed that the threshold-loss estimates will be required. The estimates obtained by Novick, Lewis, and Jackson (1972) were based on a zero-one loss function, and thus, a modification of the Novick, Lewis, and Jackson method would be desirable. At present, cumbersome numerical methods would be required to obtain a solution.

One example that rather dramatically illustrates the effect of the Bayesian estimation procedure is the following. Suppose we administer a criterion-referenced test to a group of examinees before and after instruction. Let us limit ourselves to the problem of estimating mastery scores on a particular objective for the group of examinees on the two test occasions. Suppose that the tests are short, and hence, probably have only moderate reliability. Suppose further that the mean pretest and posttest scores are .4 and .8, respectively, and the threshold score is .65. Now a student with a proportion-correct score of .7 on the pretest would under the usual procedure be allowed to skip that particular unit of instruction. However, chances are that this student's mastery score is overestimated. The Bayesian analysis might well decide that he was a nonmaster. Speaking loosely and with respect to a squared-error loss method, the

Bayesian analysis might regress his estimated score further toward the mean than the cutting score and, thus, assign him to take instruction on the skill.

Consider now a student with a proportion-correct score of .6 on the posttest. Here the Bayesian analysis could be such that his "estimated score," in effect, exceeds .65. Then, instead of assigning him to some remedial program, he will be allowed to go on to new work. However, if his posttest group had a mean performance of .68, he would probably be estimated to be a nonmaster.

Approaches to Reliability and Validity Estimation

In practical applications of criterion-referenced testing, it would seem that in order to evaluate the test, it would be necessary to know something about the consistency of decision making across parallel forms of the criterion-referenced test or across repeated measurements (i.e., reliability). Another aspect of the measurement procedure that should seemingly be considered is the accuracy of decision making (i.e., validity). The problem of reliability and validity estimation for criterion-referenced tests is considered next.

Because the designer of a criterion-referenced test has little interest in discriminating among examinees, no attempt is made to select items to produce a test of maximum test score variability, and thus, that variance will typically be small. Also, criterion-referenced tests are usually administered either immediately before or after small units of instruction. Thus, it is not surprising that we frequently observe homogeneous distributions of test scores on the pre- and posttests, but centered at the low and high ends of the achievement scales, respectively. It is well known from the study of classical test theory (Lord & Novick, 1968) that when the variance of test scores is restricted, correlational estimates of reliability and validity will be low. Thus, it seems clear that the classical approaches to reliability and validity estimation will need to be interpreted more cautiously (or discarded) in the analysis of criterion-referenced tests. Perhaps, an even more serious reservation concerning the classical approach to reliability and validity estimation for criterion-referenced tests, if one looks at these psychometric concepts in decision-theoretic terms, is that the correlational method represents an inappropriate choice of a loss function (squared-error loss in the π metric) with which to evaluate a test. This point will be expanded upon later.

However, before considering a decision-theoretic approach to reliability and validity estimation, let us review some alternate approaches proposed by other writers. Carver (1970) argues that the reliability of any test depends upon replicability, but replicability is not dependent upon test score variance. If a group of examinees all obtain similar scores (to other members of the group) on parallel forms of some criterion-referenced test, near perfect replicability exists even though test reliability, estimated using classical correlational methods, would be close to zero. This rather extreme example points out the shortcoming of the correlational approach to reliability estimation. Carver (1970) proposed two statistics to assess criterion-referenced test reliability. First, he says, "The reliability of a single form of a criterion-referenced device could be estimated by administering it to two comparable groups. The percentage that met the criterion in one group could be compared to the percentage that met the criterion in the other group [p. 56]." The more comparable the statistics, the more reliable the test could be said to be. Secondly, Carver suggested that the reliability of a criterion-referenced test should be assessed by comparing the percentage of examinees achieving the criterion on parallel tests.

Cox and Graham (1966) report the use of the coefficient of reproducibility as an alternative to the classical approach to reliability estimation for one special type of criterion-referenced test. They calculate the coefficient for a sequentially scaled criterion-referenced test designed for use in a unit of instruction where objectives can be identified as being sequential in nature. Tests are said to be scalable if for a particular ordering of items, individuals are able to answer all questions up to a point and none beyond. The coefficient of reproducibility is a measure of the extent to which

group performance satisfies this condition. As Cox (1970) suggests, the problems of using the coefficient of reproducibility as a reliability estimate have yet to be determined.

Another interesting suggestion for reliability estimation comes from the work of Livingston (1972a, 1972b). He proposes a reliability coefficient which is based on squared deviations of scores from the performance standard (or cutting score) rather than the mean as is done in the derivation of reliability for norm-referenced tests in classical test theory. The result is a reliability coefficient which has several of the important properties of a classical estimate of reliability. In fact, it can be easily shown that the classical reliability is simply a special case of the new reliability coefficient. However, several psychometricians (e.g., Harris, 1972a) have expressed doubts concerning the usefulness of Livingston's reliability estimate.

Our own feeling is that Livingston misses the point for much of criterion-referenced testing. It is not, as he suggests, "to know how far [a student's] score deviates from a fixed standard." Rather, the problem is one of deciding whether a student's true performance level is above or below some cutting score. In fact, in most practical applications of criterion-referenced tests, the test score is used to dichotomize individuals into either a "mastery" or a "nonmastery" category. Thus from our conceptualization of the measurement problem with criterion-referenced measurement, Livingston's choice of a loss function with which to evaluate the reliability of a criterion-referenced test is wrong. Specifically, we suggest that squared-error loss in the π metric *is not* appropriate and that threshold loss *is* appropriate.

Now, it may be the case that a measurement situation will arise with the new instructional models and a squared-error or absolute-error loss function may be appropriate; but in such a situation, it is unlikely that there would simultaneously be a great concern with a threshold score.

While there has been little work done on the problem of assessing reliability, even less work has been reported to date on establishing the validity of criterion-referenced test scores. Above all else, a criterion-referenced test must have content validity. According to Popham and Husek (1969), content validity is determined by "a carefully made judgment, based on the test's apparent relevance to the behaviors legitimately inferable from those delimited by the criterion." If techniques such as those advocated by Hively, Patterson, and Page (1968) or Bormuth (1970) for defining content domains and item generation rules are followed, content validity follows. If other procedures are used, the task of determining content validity becomes more difficult.

While we would suggest that the traditional concepts of reliability and validity could be replaced by a complete decision-theoretic formulation, it will nevertheless be useful to point out a relationship between these approaches. Suppose we are given two criterion-referenced tests which in a specified population and for a specified qualifying score π_0 are parallel (in the classical sense—see Lord & Novick, 1968) in the ω metric. Denote the estimates of ω for person i on the two tests by the observed scores $\hat{\omega}_{1i}$ and $\hat{\omega}_{2i}$ and define the reliability of the test as the correlation over persons of $\hat{\omega}_{1i}$ and $\hat{\omega}_{2i}$. This is, of course, classical reliability theory in the ω metric. It is not particularly satisfactory for the usual reasons that product moment correlations are unsatisfactory measures of association or agreement for binary (zero-one) variables. A more satisfactory measure of reliability might simply be the proportion of times that the same decision would be made with the two parallel instruments.

Validity theory would take the same form, except of course, that a new test Y would serve as criterion and the qualifying score on the second test need not correspond with the qualifying score on the predictor criterion-referenced test. The criterion "test" might well be derived from performance on the next unit of instruction, or it would be a job-related performance criterion.

REFERENCES

Atkinson, R. C. Computer-based instruction in initial reading. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton: Educational Testing Service, 1968.

Block, J. H. Criterion-referenced measurements: Potential. *School Review,* 1971, **69**, 289-298.

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

Bloom, B. S. Learning for mastery. *Evaluation Comment,* 1968, **1**, No. 1.

Bormuth, J. R. *On the theory of achievement test items.* Chicago: University of Chicago Press, 1970.

Carroll, J. B. A model of school learning. *Teachers College Record,* 1963, **64**, 723-733.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. *Educational Horizons,* 1970, **48**, 71-80.

Carver, R. P. Special problems in measuring change with psychometric devices. In *Evaluative Research: Strategies and Methods.* Pittsburgh: American Institutes for Research, 1970.

Cox, R. C. Evaluative aspects of criterion-referenced measurement. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970. (ERIC, ED 038 679)

Cox, R. C., & Graham, G. T. The development of a sequentially scaled achievement test. *Journal of Educational Measurement,* 1966, 3, 147-150.

Cronbach, L. J. How can instruction be adapted to individual differences? In R. M. Gagné (Ed.), *Learning and Individual Differences.* Columbus, Ohio: Charles E. Merrill, 1967.

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

Flanagan, J. C. Functional education for the seventies. *Phi Delta Kappan,* 1967, **49**, 27-32.

Flanagan, J. C. Program for learning in accordance with needs. *Psychology in the Schools,* 1969, **6**, 133-136.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist,* 1963, **18**, 519-521.

Glaser, R. Adapting the elementary school curriculum to individual performance. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton: Educational Testing Service, 1968.

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

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

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement,* 1972, **9**, 27-29. (a)

Harris, C. W. An index of efficiency for fixed length mastery tests. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972. (b)

Harris, M. L., & Stewart, D. M. Application of classical strategies to criterion-referenced test construction. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hieronymus, A. N. Today's testing: What do we know how to do? In *Proceedings of the 1971 Invitational Conference on Testing Problems.* Princeton: Educational Testing Service, 1972.

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

Ivens, S. H. An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Jackson, P. H., & Novick, M. R. Maximizing the validity of a unit-weight composite as a function of relative component lengths with a fixed total testing time. *Psychometrika,* 1970, **35**, 333-347.

Kriewall, T. E. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Livingston, S. A. Criterion-referenced applications of classical test theory, *Journal of Educational Measurement,* 1972, **9**, 13-26. (a)

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." *Journal of Educational Measurement,* 1972, **9**, 31. (b)

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. *Psychometrika,* 1972, **37**, in press.

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

Suppes, P. The uses of computers in education. *Scientific American,* 1966, **215**, 206-221.

Woodbury, M. A., & Novick, M. R. Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time. *Journal of Mathematical Psychology,* 1968, **5**, 242-259.

ACT Research Reports

This report is Number 53 in a series published by the Research and Development Division of The American College Testing Program. The first 26 research reports have been deposited with the American Documentation Institute, ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D.C. 20540. Photocopies and 35 mm. microfilms are available at cost from ADI; order by ADI Document number. Advance payment is required. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress. Beginning with Research Report No. 27, the reports have been deposited with the National Auxiliary Publications Service of the American Society for Information Science (NAPS), c/o CCM Information Sciences, Inc., 22 West 34th Street, New York, New York 10001. Photocopies and 35 mm. microfilms are available at cost from NAPS. Order by NAPS Document number. Advance payment is required. Printed copies ($1.00) may be obtained, if available, from the Publication and Information Services Division, The American College Testing Program, P.O. Box 168, Iowa City, Iowa 52240. A check or money order must accompany the request.

The reports since May 1970 in this series are listed below. A complete list of the reports can be obtained by writing to the Publication and Information Services Division, The American College Testing Program, P. O. Box 168, Iowa City, Iowa 52240.

No. 34 *Research Strategies in Studying College Impact,* by K. A. Feldman (NAPS No. 01211; photo, $5.00; microfilm, $2.00)

No. 35 *An Analysis of Spatial Configuration and Its Application to Research in Higher Education,* by N. S. Cole, & J. W. L. Cole (NAPS No. 01212; photo, $5.00; microfilm, $2.00)

No. 36 *Influence of Financial Need on the Vocational Development of College Students,* by A. R. Vander Well (NAPS No. 01440; photo, $5.20; microfilm, $2.00)

No. 37 *Practices and Outcomes of Vocational-Technical Education in Technical and Community Colleges,* by T. G. Gartland, & J. F. Carmody (NAPS No. 01441; photo, $6.80; microfilm, $2.00)

No. 38 *Bayesian Considerations in Educational Information Systems,* by M. R. Novick (NAPS No. 01442; photo, $5.00; microfilm, $2.00)

No. 39 *Interactive Effects of Achievement Orientation and Teaching Style on Academic Achievement,* by G. Domino (NAPS No. 01443; photo, $5.00; microfilm, $2.00)

No. 40 *An Analysis of the Structure of Vocational Interests,* by N. S. Cole, & G. R. Hanson (NAPS No. 01444; photo, $5.00; microfilm, $2.00)

No. 41 *How Do Community College Transfer and Occupational Students Differ?* by E. J. Brue, H. B. Engen, & E. J. Maxey (NAPS No. 01445; photo, $5.50; microfilm, $2.00)

No. 42 *Applications of Bayesian Methods to the Prediction of Educational Performance,* by M. R. Novick, P. H. Jackson, D. T. Thayer, & N. S. Cole (NAPS No. 01544; photo, $5.00; microfilm, $2.00)

No. 43 *Toward More Equitable Distribution of College Student Aid Funds: Problems in Assessing Student Financial Need,* by M. D. Orwig (NAPS No. 01543; photo, $5.00; microfilm, $2.00)

No. 44 *Converting Test Data to Counseling Information,* by D. J. Prediger (NAPS No. 01776; photo, $5.00; microfiche, $2.00)

No. 45 *The Accuracy of Self-Report Information Collected on the ACT Test Battery: High School Grades and Items of Nonacademic Achievement,* by E. J. Maxey, & V. J. Ormsby (NAPS No. 01777; photo, $5.00; microfiche, $2.00)

No. 46 *Correlates of Student Interest in Social Issues,* by R. H. Fenske, & J. F. Carmody (NAPS No. 01778; photo, $5.00; microfiche, $2.00)

No. 47 *The Impact of College on Students' Competence to Function in a Learning Society,* by M. H. Walizer, & R. E. Herriott (NAPS No. 01779; photo, $5.00; microfiche, $2.00)

No. 48 *Enrollment Projection Models for Institutional Planning,* by M. D. Orwig, P. K. Jones, & O. T. Lenning (NAPS No. 01780; photo, $5.00; microfiche, $2.00)

No. 49 *On Measuring the Vocational Interests of Women,* by N. S. Cole (NAPS No. not available at this time.)

No. 50 *Stages in the Development of a Black Identity,* by W. S. Hall, R. Freedle, & W. E. Cross, Jr. (NAPS No. not available at this time.)

No. 51 *Bias in Selection,* by N. S. Cole (NAPS No. not available at this time.)

No. 52 *Changes in Goals, Plans, and Background Characteristics of College-Bound High School Students,* by J. F. Carmody, R. H. Fenske, & C. S. Scott (NAPS No. not available at this time.)