DOCUMENT RESUME

ED 041 052 24 TM 000 023

AUTHOR Jackson, Rex
TITLE Developing Criterion-Referenced Tests.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
PUB DATE Jun 70
CONTRACT OEC-0-70-3797 (519)
NOTE 18p.

EDRS PRICE EDRS Price MF-$0.25 HC-$1.00
DESCRIPTORS *Criterion Referenced Tests, Item Analysis, *Measurement Techniques, Norm Referenced Tests, *Test Construction, Test Interpretation, Test Reliability, *Tests, Test Validity

ABSTRACT

Present definitions of the criterion-referenced test are discussed, insufficiences noted, and a new definition proposed. Some examples of criterion-referenced tests are examined and used to educe some general principles for the development of such tests. The utility of item form processes is assessed. It is suggested that the difficulty of objectively defining a test construction process is directly proportional to the complexity of the behavior the test is designed to assess. Problems and doubts with regard to the development of criterion-referenced tests for complex behavior domains are noted. In addition, some empirical methods for dealing with item analysis, test reliability, and test validity difficulties are advanced. (DG)

DEVELOPING CRITERION-REFERENCED TESTS

Rex Jackson
Test Development Division
Educational Testing Service, Princeton, N.J.

June 1970

ERIC Clearinghouse on Tests, Measurement, & Evaluation

According to Wang (1969), a "'criterion-referenced test' is an achievement test developed to assess the presence or absence of a specific criterion behavior described in an instructional objective." The term criterion-referenced appears to have been introduced by Glaser (1963) in a paper in which he distinguishes "criterion-referenced" from "norm-referenced" testing. In the latter, an individual's test performance is interpreted with respect to the performance of other individuals who belong to some specified population. In contrast, the interpretation of an individual's performance on a criterion-referenced test is a behavioral statement (or set of such statements) that is made without reference to the performance of other individuals.

Although the term "criterion-referenced" has been introduced rather recently, it should be recognized that the types of interpretations that the term implies are not a sudden innovation. Indeed, such interpretations have probably been the rule rather than the exception. Ebel (1970) points out that the percentage-mastery grades once widely favored in schools and colleges in this country represent one type of criterion-referenced measurement (albeit one that is generally unsatisfactory in practice). Many work-sample and performance measures used in personnel selection and evaluation represent another type of criterion-referenced measurement. Even when tests are described as norm-referenced, test users are frequently interested in determining what an individual can or cannot do, rather than his standing in some given population. Angoff (1962) in a discussion of scales with "non-meaningful origins and units of measurement" (i.e., scales without inherent normative meaning) points out that such scales can "derive meaning from the experience that the user acquires in applying the scale to the measurement of familiar objects." The same is

true of course of normative scales provided that stability of the scale is maintained over a sufficient length of time.

Norm-referenced measures appear to be particularly relevant in certain special decision situations--such as fixed quota selection. Glaser (1963) contends that the emphasis on norm-referenced tests in education "has been brought about by the preoccupation of test theory with aptitude, and with selection and prediction problems." Even in these special decision situations, however, reference is generally made to some behavioral criterion. Angoff (in press) points out that an expectancy table relating, say, scores on an achievement test to course grades provides meaning to test scores in terms already familiar to test users. Similarly, one may view regression methods as a way of expressing test score scales with reference to some ultimate criterion of performance.

Any test samples the content of some specified domain. Even though a test may be normed so that an individual's score may be compared with scores of some specified group, there is the assumption of some latent trait upon which observed scores depend, and which the test is, therefore, said to measure. Hence, there is always an implicit behavioral element, and even tests that are described as norm-referenced are designed to yield inferences about, say, the amount of trait X that an individual has. In contrast to a criterion-referenced test, however, the inference is of the form--"more (or less) of trait X than the mean amount in population Y"--rather than some specified amount that is meaningful in isolation.

The foregoing discussion is intended to suggest that the definitions of the term "criterion-referenced" offered by Wang and Glaser are not sufficient

to differentiate criterion-referenced and norm-referenced tests, since even tests which are described as norm-referenced may yield scores which may also be compared with a performance standard. In order to further distinguish what is meant by a criterion-referenced test, the term "criterion-referenced" will be used here to apply only to a test designed and constructed in a manner that defines explicit rules linking patterns of test performance to behavioral referents. For a norm-referenced test, the performance of some reference population is used to provide a measure of statistical control for lack of standardization in the test construction process. A criterion-referenced test, on the other hand, must be produced by an "objectively defined process." The meaningfulness and reproducibility of test scores derives then from the complete specification of "the operations used to measure the quantity involved" (Ebel, 1962).

It should be noted that the definition of criterion-referenced used here excludes tests for which test scores have simply been equated to some scale of attainment. This kind of approach has sometimes been used in order to allow functional interpretations of test scores. Tests of this type do not necessarily differ from conventional norm-referenced tests in any fundamental way; it is simply provision of an interpretive device that allows functional interpretations to be made. Although this approach will be exemplified by some of the tests discussed later in this paper, it should be recalled that the fact that content standards have been associated with test scores is not sufficient to distinguish the test as a criterion-referenced test.

Developing Criterion-Referenced Tests--
General Principles and Examples in Reading

Ebel (1962) describes the development of a criterion-referenced test of knowledge of word meanings as follows:

> Parallel forms of the test were produced, one by a test specialist and the other by an intelligent secretary who had no special training in test construction. Both tests were built on the basis of detailed written specifications and directions. The tests were based on a spaced sample of 100 words from a specified dictionary. Explicit instructions were given for choosing a unique but representative sample, and for limiting the sample to words appropriate for the test. For each word the first synonym or defining phrase was copied from the dictionary. The words were arranged in alphabetical order in a single list. The defining phrases were also placed in alphabetical order and numbered from 1 to 100. The student's task was to match the definitional phrase with the appropriate word.

The following characteristics can be distinguished in the development of Ebel's test:

1 - Specification of the universe to which generalization is desired

2 - A systematic plan for sampling from the universe

3 - A standardized method of item development (objective process for determining both correct and incorrect alternatives).

These characteristics together then serve to define the meaning of test scores. It is perhaps arguable whether such a test provides a useful estimate

of the fraction of words in the given dictionary for which a given individual knows the meaning. Perhaps other procedures would yield better estimates. Nevertheless, the procedure used was an objective one. To the extent that scores are reproducible on tests developed independently under the same procedures, the scores may be said to have inherent meaning.

Flanagan (1962) indicates that a variant of the procedure described by Ebel was used in Project TALENT. The tests designed for operational use in the areas of spelling, vocabulary, and reading were not criterion-referenced. Instead special tests were developed in these areas by systematically sampling relevant domains. For the special spelling test a list of 5000 frequently used words was sampled; for the special vocabulary test, a group of dictionaries was used; for the special reading test, passages from authors judged to represent various levels of difficulty were selected. For each of the areas, then, the regular tests were equated to the special tests so that interpretations could be made in terms of content.

The Reading Vocabulary and Reading Comprehension tests in the Progressive Achievement Tests developed by the New Zealand Council on Educational Research exhibit both of the approaches discussed above. The planning of these tests is described in Elley (1967). Further information on these tests was gained through personal communication with Dr. Elley.

The words tested on the Vocabulary test were randomly selected from the Wright List of 10,000 words--a list of words with associated frequency counts in written English that is widely used in the New Zealand schools. Synonym type items were then used to test the sampled words. In some respect, the development of items was not objective. Although certain item generation

rules were used, item writers still had latitude in choosing among alternative correct or incorrect responses. In addition, no synonyms were found for certain words selected and these words were not tested. These nonsystematic aspects of the construction of the test were due in part to the fact that the test was designed for a variety of uses. They illustrate, however, some of the difficulties of developing criterion-referenced tests in practical settings. Can systematic procedures for developing items which measure intended objectives be developed? For multiple-choice tests, how can the effects on test performance due to subjective distracter selection be controlled? Do "item writer effects" bias the process or do they balance out in a large sample of items? These issues will be discussed in more detail later in this paper.

The methods used to develop content standards for the New Zealand Reading Comprehension Test were similar to the Project TALENT methods discussed above. Passages included in the test were rated for difficulty by means of a readability formula involving noun frequency counts (Elley, 1969). Passages were presented in each form of the test in order of difficulty and conventional item analyses were used to insure that the sets of questions based on the passages were in the same order of difficulty as the passages themselves. Three experiments were then conducted in which students were asked to read the passages for one form of the test aloud and answer five questions on each passage. A student was stopped when he failed to answer three out of five questions correctly. By relating scores on another form of the test to the difficulty ratings of the last passages attempted by students, it was possible to develop a table indicating what kind of reading material is suitable for a student achieving a given score. Again, the process of test development was

not entirely objective. A test developed under identical procedures but including a different sample of passages and questions might well lead to inferences differing systematically from those provided by the actual test. Indeed it appears that Elley's scaling experiments were required, because the process of the test construction could not be specified in sufficient detail.

The brief accounts of test development described above suggest the following generalization: the difficulty of objectively defining a test construction process is directly related to the complexity of the behavior the test is designed to assess. For tests of spelling and knowledge of word meanings it is possible to specify finite universes of content (word lists) from which random samples could be drawn. In addition, for one of the vocabulary tests discussed, an objective method for developing test items was used. A method for systematically generating spelling items is described in Fremer and Anastasio (1969). On the other hand, for tests of reading comprehension, it was not clear that similar procedures could be used, and the question remains whether suitable criterion-referenced test construction procedures can be found that will allow generalization from test performance to complex behavior domains.

Osburn (1968) discusses two conditions that he sees as prerequisites for allowing inferences to be made about a domain of knowledge from performance on a collection of items: "The first is that all items that could possibly appear in the test should be specified in advance. Secondly, the items in a particular test should be selected by random sampling or stratified random sampling from the universe of content." The first of these conditions is generally difficult to satisfy in a reasonable way for complex domains.

However, the problem of listing the elements of a universe of item content can be overcome to a degree, if a generative process can be defined which could in theory produce such a listing. One such process has been termed an "item form." Osburn has described the item form process in terms of the following characteristics: "(1) it generates items with a fixed syntactical structure; (2) it contains one or more variable elements; and (3) it defines a class of item sentences by specifying the replacement sets for the variable elements." Osburn goes on to point out that a distinguishing characteristic of the item form method is that there exists an "unbroken link" between the generative system and the specific item produced. A collection of item forms (perhaps a hierarchically ordered one) together with the replacement sets for the variable elements then defines a universe of content.

In practice, an item form consists of a sentence with one or more blanks. The words that fit in the blanks may be systematically varied to produce items of different levels of specificity. Since the procedure is systematic and rule bound, it has proved amenable to automation. Shoemaker and Osburn (1969) have reported the construction of a computer program, "capable of generating random or stratified random parallel tests from a specified content population." The input to the program consists of a sentence frame like the following:

> Given a normal distribution with mean equal to___and standard deviation equal to____. If one number is randomly sampled from this distribution, what is the probability that this number will be greater than or equal to____?

A random number generator is used to supply values for the blanks in the item form. It is possible to specify acceptable ranges of replacement numbers in order to create realistic problems or reduce computation. The program is not

limited to numeric substitution; it may randomly choose replacement elements from any list that is supplied to it.

The concept of item forms has much in common with the "facet design" method advocated by Guttman (Guttman & Schlesinger, 1966), in which a sentence frame is produced with a number of variable elements called "facets." An attractive element of the Guttman procedure is the systematic generation of distracters for multiple-choice items. A rather trivial example follows:

A puppy is a(n) __x__ __y__ .

Where x could have the values (young, old) and y could have the values (dog, cat, cow, pig, horse,...).

It should be noted that the major goal of these methods is to allow inference from test performance to behavioral referents. By these methods, items can be completely specified according to rules. As discussed previously, a major advantage lies in the ability to produce groups of items that are random or stratified random samples of a specified universe of content. It should be noted, however, that the generalization afforded by these methods is to the particular universe of content defined by the sentence frames and variable elements. Interpretations of test performance in terms of such constructs as "reading comprehension" or "mathematical ability" are not themselves made legitimate by the process of test construction. This is to say, functional interpretations of test performance are (in the absence of other evidence) necessarily limited to the rather narrow processes embodied in the item generating rules. It is important to note that the fact that an objective process of test construction can be defined does not necessarily mean that test performance can be interpreted in terms of some theory that gave rise to the item generating rules.

Empirical Evaluation of Criterion-Referenced Items and Tests

Popham and Husek (1969) cast doubt on the applicability to criterion-referenced tests of conventional item analysis procedures and methods of assessing reliability and validity. The burden of their argument is that scores on a criterion-referenced test may have no variance in some population of interest, and yet the test may be a good test. That is to say, this lack of variance does not necessarily imply that the test is ineffective for the purposes for which it was designed. Indeed on certain criterion-referenced tests, it is possible that all students completing a particular course of instruction will pass every item (see Cartier, 1968).

With respect to selection of items on the basis of item analysis data, Osburn makes the following observation:

> It is evident that these procedures may bias the inferences regarding a person's true score on the universe of content, and the nature of the bias will generally be unknown... Rejection of the item always implies rejection of the class of items to which the item belongs or at least a modification of the generating rule that specifies the item class.

It is clear that Osburn's remarks apply to criterion-referenced tests as defined in this paper. When comparability of test scores and behavioral standards is postulated upon systematic sampling of tasks from a universe of content, it is difficult to see how item selection could legitimately be influenced by item analysis data.

In certain respects, however, item analysis may have value. Cox and Vargas (1966) investigated a number of different discrimination indices

including one considered particularly relevant to criterion-referenced tests--an index of an item's ability to discriminate pre- and post-training performance. A post-hoc analysis of this type could have value in assisting evaluation of both the relevance of the test and the adequacy of the intervening training. If a negatively discriminating item does in fact reflect accurately an instructional objective, the effectiveness of the instruction must be questioned.

Although Ccx and Vargas discussed discrimination of pre- and post-training groups, it is clear that any two contrasted groups could be used as long as one group is known to have mastered the behavior in question to a greater degree than the other. If items belonging to a certain class are consistently poor at discriminating such contrasted groups, it would be necessary to revise the model of the behavioral domain under which the test was developed. Item analyses of this type, therefore, can serve as an empirical check on the validity of the hypothetical constructs the test is intended to measure.

With respect to reliability, the possibility that scores on a criterion-referenced test may have no variance for some population of interest does cast doubt on the relevance of the concept of reliability as defined in classical test theory. In most practical settings, however, it seems likely that the performance of individuals taking criterion-referenced tests will vary. When such a test does yield a metric scale, it appears that conventional internal consistency and correlational methods for estimating reliability can be used. If a set of criterion-referenced tests are, in fact, random samples from a well defined domain, Cronbach's Generalizability Theory may be applied in estimating components of variance due to various error sources (Cronbach et al., 1963). Application of this theory to a group of arithmetic tests is described by Hively, Patterson, and Page (1968).

As a more general principle, however, it appears that criterion-referenced analogues to the traditional concept of reliability are needed. In certain cases, the ultimate "reported scores" for a criterion-referenced test may be only nominal or ordinal in character. One way that "reliability" might be analyzed is through comparison of the inferences made for a group of individuals on one form of a test with the inferences yielded by an alternate form developed independently with identical procedures. An index of agreement between the two forms in classifying the individuals tested--perhaps a contingency coefficient--could then be used as an index of the "reliability" of the measurement procedure. It may be noted that this form of reliability estimation is in some respects more rigorous than correlational methods (for tests with metric scales) which are unaffected by changes in the origin or unit of measurement.

With respect to validity, Osburn (1968) makes the following point:

> What the test is measuring is operationally defined by the universe of content as embodied in the item generating rules. No recourse to response-inferred concepts such as construct validity, predictive validity, underlying factor structure or latent variables is necessary to answer this vital question.

What Osburn is discussing may be termed the definitional validity of a criterion-referenced test. If test scores are interpreted solely in terms of measurement operations, the interpretations are tautologically valid. In this context, the alternate forms types of experiment might be considered as a check on validity. Such an experiment would in effect provide cross-validation of the representativeness of the particular samples of tasks in each of the tests. Ebel (1962) describes such an experiment for the vocabulary

test described in an earlier section of this paper. In essence, "reliability" (stability of results across samples of items) is considered a sufficient, rather than only a necessary condition for validity. If observed scores do provide reliable estimates of an individual's "true score" on a universe of content, they are ipso facto valid for that universe.

It may be noted that a rather narrow concept of validity has been used above. In many cases, generalization beyond the universe of content defined by the item generating rules will be desired. Frequently, it will be convenient to interpret scores in terms of some hypothesized psychological construct or to use scores to predict performance on tasks of types different from those included in the test. In order to validate these interpretations of test performance, conventional methods of assessing validity could be used (i.e., either by direct correlational methods for observable variables or by testing of hypotheses suggested by the theories underlying constructs). It is clear that interpretations of this type would be made only when the relevant individuals exhibit variation on the dependent variable. When this is the case, no test variance would indeed suggest that the interpretations offered are invalid.

One further procedure for empirically evaluating certain criterion-referenced tests deserves mention. For tests designed to indicate what set of tasks an individual has mastered within a hierarchical or otherwise structured domain, it is a considerable convenience if the response pattern of any individual can be inferred from a single score (or perhaps from a limited set of scores)--that is if the items in the test form a Guttman scale. Methods have been developed for determining the "scalability" of tests in this sense and the smallest variable space which satisfactorily maintains the

information in the original item responses. Cox and Graham (1966) report an attempt to develop a short elementary arithmetic test yielding a unidimensional Guttman scale. However, many measurement specialists believe that Guttman scales can be attained for achievement tests only in trivial cases.

Conclusion

Interest in criterion-referenced tests has risen in recent years as it has become increasingly clear that measures allowing only population-referenced interpretations do not provide the information that is needed in making certain types of decisions in education. Criterion-referenced measures have been considered particularly desirable in areas where diagnostic information is needed, such as placement of individuals in programs of instruction or individual instruction, in formative evaluation of educational programs, and in evaluative assessment of individual or group achievement. There is some doubt, however, that suitable, "pure" criterion-referenced tests can be developed for complex domains. Ebel (1970) in a paper on the limitations of criterion-referenced tests has articulated what is undoubtedly a widely held belief: "Criterion-referenced measurement may be practical in those few areas of achievement which focus on cultivation of a high degree of skill in the exercise of a limited number of abilities." It appears that at the current state of the art, it is difficult to develop the objective procedures necessary for criterion-referenced measurement of complex behavior without doing violence to measurement objectives. What is needed for complex domains are item generating rules that permit generalizations of practical significance to be made.

The above is not intended to imply that attempts to relate test performance to behavioral statements are pointless. It was noted earlier that experience

in applying a norm-referenced scale to familiar objects and experimental evidence linking such a scale to performance criteria provide some basis for making functional interpretations of test performance. For complex behavior domains, it appears that at least until explicit models stated in measurable terms are developed, a degree of subjectivity in test construction (and attendant population-referenced scaling) will be required. Both detailed specification (though not complete standardization) of test construction processes and experimental evidence relating behavior to test and item performance appear to be the most promising approach for at least the near future.

REFERENCES

Angoff, W. H. Scales with nonmeaningful origins and units of measurement. Educational and Psychological Measurement, 1962, 22, 27-34.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement, in press.

Cartier, F. A. Criterion-referenced testing of language skills. Tesol Quarterly, 1968, 2. Abstracted in Research in Education, Educational Resources Information Center, (ED 020 515).

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

Cox, R., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Report number BR-5-0253. Learning Research and Development Center, University of Pittsburgh, February 1966. Abstracted in Research in Education, Educational Resources Information Center, (ED 010 517).

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Ebel, R. Content standard scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, R. Some limitations of criterion-referenced measurement. Paper prepared for AERA Symposium "Criterion-referenced measurement: Emerging issues," Minneapolis, March 1970.

Elley, W. B. Standardized test development in New Zealand: Problems, procedures, and proposals. New Zealand Journal of Educational Studies, 1967, 2, 63-77.

Elley, W. B. The assessment of readability by noun frequency counts. Reading Research Quarterly, 1969, 4, 411-427.

Flanagan, J. C. Discussion of symposium: Standard scores for aptitude and achievement tests. Educational and Psychological Measurement, 1962, 22, 35-39.

Fremer, J., & Anastasio, E. Computer-assisted item writing-I (Spelling Items). Journal of Educational Measurement, 1969, 6, 69-74.

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

Guttman, L., & Schlesinger, I. M. Development of diagnostic analytical and mechanical ability tests through facet design and analysis. The Israel Institute of Applied Social Research, Jerusalem, Israel, 1966.

Hively, Patterson, & Page. A universe-defined system of arithmetic tests. Journal of Educational Measurement, 1968, 5, 275-295.

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.

Shoemaker, D. M., & Osburn, H. G. Computer-aided item sampling for achievement testing. Educational and Psychological Measurement, 1969, 29, 165-172.

Wang, M. C. Approaches to the validation of learning hierarchies. Western Regional Conference on Testing Problems (Proceedings) 1969. Princeton, N. J.: Educational Testing Service, 14-38.