DOCUMENT RESUME

ED 125 577 IR 003 661

AUTHOR Swezey, Robert W.
TITLE Toward the Development of Realistic Measures of Performance Effectiveness.
PUB DATE [76]
NOTE 22p.; Paper presented at the International Learning Technology Congress and Exposition on Applied Learning Technology for Human Resource Development (Washington, D.C., July 21-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Speeches; Test Construction; Testing

ABSTRACT

Though domain-oriented and norm-referenced tests are appropriate for some situations, objective-oriented and criterion-referenced tests must be used to gather additional information. Objectives for such tests must include a statement of the desired performance, the test conditions, and the standards of acceptance. When tests are constructed the following questions should be considered: (1) fidelity--the degree to which the test resembles the desired outcome; (2) objectivity of the measurement; (3) scoring problems; (4) emphasis on product or process; (5) reliability; and (6) content, concurrent, and predictive validity. (EMH)

Toward the Development of Realistic Measures of Performance Effectiveness

Robert W. Swezey

Applied Science Associates

Reston, Virginia

Toward the Development of Realistic Measures of Performance Effectiveness

The topic about which I would like to comment today, concerns methods for evaluating the effectiveness of individual performance. The individual of whom I speak, might for example, be an electronics technician on a radar system, a medical student, a second grader, or a pianist. The context in which the individual functions; while important, and most certainly a variable in the equation which describes performance effectiveness; is not the item of interest here. Instead, I would like to discuss some generalized topics concerned with the assessment of individual performance. Notice that I speak of performance, not of attitudes, knowledge, abilities, or other so-called intervening variables. I believe that we should be concerned with performance outcomes; with actions or statements which we can observe, define, and measure. Although this insistence may be viewed as severely restricting the applicability of certain measurement models, ... so be it. I do not wish to speculate about the effects of personality, attitudes and the like, on performance. While these are interesting and possibly productive areas to pursue, they are not the topic of concern here.

Before proceeding, let me introduce some terms. These terms are descriptive ones, which define various models of performance (and other) testing. I would like to review them briefly.

The most widely used model for assessing individual achievement is generally referred to as Norm-Referenced Measurement (NRM). In NRM, the performance of an individual is typically considered relative to the performance of other comparable individuals. This model has benefited from many years of psychometric research. It is useful for making decisions among individual attainment, and for comparing individuals to normative distributions. It allows for the possibility of ranking persons according to competence on specific tasks, or on more general measures of achievement. In cases where relative decisions must be made; such as selection, promotion, pay level judgments, class rankings, and other discriminations among individuals, NRM is the model of choice.

For example, if we have a test where local norms have been computed over a period of time, and we discover that an individual's test score is at the ninetieth percentile of that distribution, we may conclude that the person of interest is doing better than about ninety percent of the individuals in the population. A key emphasis of norm-referenced measurement, is to maximize individual differences so that one can spread the distribution of test scores. Norm-referenced items thus, are designed to discriminate, and are often chosen to be of moderate or extreme difficulty.

Unfortunately, however necessary NRM may be in performance evaluation systems, it is not sufficient. Many educational institutions are finding themselves in the position where minimal required levels

of competence are not being met. It is widely alleged, for example, that many high school graduates cannot read acceptably. To the extent that this (and similar allegations) are correct, they are difficult to detect with a norm-referenced model. The reason is that absolute performance standards are not specified in NRM. No external criterion exists, against which to assess individual performance. A different measurement model, termed Criterion-Referenced Measurement (CRM), is appropriate. A criterion-referenced test measures what an individual can do, or knows, compared to what he must be able to do, or must know, in order to successfully complete a task. Basically, this means that an individual's performance is compared, or referenced, to some external criterion, or performance standard. Such standards are derived directly from an analysis of what is required to perform a particular task successfully. In CRM, performance is interpreted against an absolute standard without regard to the distribution of scores attained by other individuals.

The distinction between NRM and CRM has been aptly illustrated by Popham and Husek (1969) using the analogy of a dog owner who wants to keep his dog in the back yard. The owner finds out how high the dog can jump (a criterion-referenced test) and builds a fence high enough to keep the dog in the back yard. How high the dog can jump compared to other dogs (a norm-referenced test) is irrelevant. Beginning with Glaser (1963) a number of researchers have made similar distinctions. Glaser and Nitko (1971, p. 653) for example, have

described a criterion-referenced test as "One that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards." This definition has been slightly expanded by Livingston (1972, p. 13) "Criterion-referenced (is) used to refer to any test for which a criterion score is specified without reference to the distribution of scores of a group of examinees." Common to all definitions is the notion that a well-defined content area and the development of procedures for generating appropriate samples of test items are important.

Two other models of performance specification will also be mentioned. Domain-Referenced Measurement (DRM), has been defined by Sanders and Murray (1976) as "a test in which performance on a task in interpreted by referencing a well-defined set of tasks (a domain)." Domain-referenced tests thus, are tests which emphasize the creation of item pools or item forms, representative of a universe of all test items for a well-defined content area.

Another model, Objectives-Referenced Measurement (ORM), is generally considered as measurement in which performance is interpreted by referencing the behavioral objective(s) for which the item was written. Objectives-referenced tests emphasize test items which are derived directly from predetermined behaviors. ORTs thus, are tests whose items are operational definitions of behavioral objectives. (See Sanders and Murray, 1976, for a further discussion of these topics.)

It appears thus, that domain- and objectives-referenced measurement refer generally to the content which the test was developed to assess. Norm- and criterion-referenced measurement, on the other hand, refer generally to the way in which a test score is interpreted, regardless of content.

Many sophisticated models for the development and validation of achievement measures exist. The problem with many of these models in every day situations, is that their esoteric nature and complicated procedures often serve to minimize their utility. Classroom teachers, it is alleged, rarely consider questions of reliability and validity in their test development activities. One reason for this may be that the establishment of test reliability and validity generally involves complicated procedures, as well as a great deal of work. It is often neither cost-effective nor time-effective for a public school teacher to compute item statistics or test reliability and validity coefficients. A typical approach to test development in applied educational contexts, is simply: (1) to determine the domain which one wants to test, (2) to write a number of test items relevant to that domain, (3) to administer the test to the appropriate student population, (4) to score the test as objectively and unbiasedly as possible, and (5) to arbitrarily establish cutting points for the grade distribution. It is here suggested that this may be a reasonable approach if one's purpose in developing the test, is a norm- and/or domain-referenced

one. If however, one is concerned about objectives, and criterion-referenced measurement (and I firmly believe that we must be concerned about these aspects) the approach is generally inappropriate.

It is not my purpose today to go into all diverse components of individual performance measurement systems, but to describe briefly certain aspects of objective-oriented, criterion-referenced systems which I believe to be of general interest. The areas about which I would like to comment, are often considered to be troublesome ones. They have generated a great deal of discussion and comment, yet so far as I know, there exists today no general agreement concerning their solution.

Objectives. First, let us consider behavioral objectives. It is my belief that adequate behavioral objectives can, generally speaking, be divided into three components. These are: performances, conditions and standards.

Performances. Every objective should state precisely what the individual must do. The statement of performance must be clear enough for that performance to be trained and tested. Examples of adequate performance statements are: climb the telephone pole; state the conditions under which a tourniquet should be applied; add two 5-digit numbers, ... etc. Every statement of performance should include an action verb. This verb is the key to the performance. It tells what must be done. In the example "state the conditions under which a tourniquet should be applied," the action verb is "state". You can actually test a student's ability to state the required conditions.

Suppose that the statement of performance had read, "appreciate the situations in which a tourniquet should be applied." Would you know what to test? How would you know when a student appreciates situations?

Conditions. Every objective should also include a statement of the conditions under which the performance must be demonstrated. Such statements should indicate: (a) what the student has to work with (or what he is allowed to use), (b) the circumstances under which the performance must be demonstrated, (c) what the student must work on (his starting points), and (d) limitations or special instructions. It is extremely important for an objective to specify all conditions which may affect performance. Without statements of the conditions, one cannot be sure of what to teach or test. Suppose, for example, that an objective stated, "compute the square root of the number 125." You, the student, have received training in the computation of square roots, and are ready to be tested. An unknown examiner takes you to a room, closes the door, and asks you to compute the appropriate square root. Your response ... "But, during my training in square root computation, I had access to a calculator." The examiner's answer, "It is important to be able to compute square roots under any circumstance; you won't always have a calculator."

The point of this rather simple example is that, if conditions aren't specified, the student won't know exactly what he needs to learn to do, and the test developer won't know just what it is he should test. A precise specification of the conditions under which the performance must be demonstrated is critical.

Standards. Thirdly, each objective should specify precisely the standard or criterion against which performance is to be evaluated. As is the case for statements of performance and conditions, standards too must be clearly stated in the objective. For an example, suppose that an objective stated, "Be able to type accurately using an electric typewriter under standard office conditions." Lacking standards for speed and accuracy, how fast would you train people to type in order to satisfy the objective? How fast would they have to type to be able to pass your criterion-referenced test? Obviously, the statement is lacking a clear statement of standards. "Accurately" doesn't really tell you anything. A complete objective might read: "Using an electric typewriter under standard office conditions, be able to type 50 words per minute, corrected for accuracy (that is, one word subtracted for each mistake). Working from such an objective, you would know what standards to shoot for in training, and the level of performance the examinee must demonstrate on the test.

A final comment on objectives, is that they must be unitary. They should cover one task or task aspect only. To check that objectives are unitary, one should examine the parts that describe the performance. Looking at the performance required by a given objective, one might ask oneself the following two questions: (a) Does the objective call for performance on just one task? (b) Are all tasks independent (that is, success on one objective does not require successful performance on the preceding one)? If the answer to either

question is a definite "no", the objectives are probably not unitary, and need to be broken down into unitary ones.

Item Format and Level of Fidelity. The second topic which I wish to comment upon, concerns item format and level of fidelity. Before constructing test items, the developer is typically faced with questions of item format. Do we want paper and pencil items, "hands-on" performance items, multiple choice items, recall measures, job simulation, supervisor ratings, or what? Virtually any of these formats can be adapted to a testing situation. There may be others that are even more appropriate. How to choose? These are questions involving item format and test fidelity.

The term fidelity addresses the extent to which a test resembles the actual objective or performance being examined. The more the test resembles the performance in question, the higher the fidelity of the test. Here is one place where practical testing constraints have a direct impact on test development. If, for example, it is too closely to use an actual aircraft for maintenance tests, and one must therefore use a simulator, one loses fidelity unless the simulator is very much like the actual aircraft in terms of required performance. To the extent that the performances required on the simulator approach those required on the actual equipment, the fidelity loss is minimized.

Frederiksen (1962) has proposed a multiple level classification of fidelity in performance testing. The first category (and lowest fidelity level) is to solicit opinions. This category may in fact often

miss the payoff question (e.g., to what extent has the behavior of trainees been modified as a function of the instructional process). The second category is to administer attitude scales. This technique, although psychometrically refined via the work of Thurstone, Likert, Guttman and others, assesses primarily a psychological concept (attitude) which is presumed to be concomitant with performance. Third is to measure knowledge. This is without doubt, the most commonly used method of assessing achievement. This technique is usually considered adequate however, only if the training objective is to produce knowledge. Fourth: elicit related behavior. This approach is often used in situations where, due to practical considerations, one must resort to observation of behavior which is thought to be logically related to the criterion behavior. Fifth: elicit "What I Would Do" behavior. This technique usually involves the presentation of brief descriptions of problem situations or scenarios, under simulated predesigned conditions, and requires a subject to indicate what he would do to solve the problems if he were in the situation. And finally, at the highest fidelity level--elicit lifelike behavior. This category includes behavioral assessment under conditions which approach the realism of the life situation. Flight simulators for example, fall into this category.

A good guideline for item format, is that the item should be in the form that best approximates the behavior specified by the objective. If the instruction is aimed at problem solving, for instance,

then the items should address problem solving tasks and not, for example, knowledge about required background content. If the instruction is intended to evaluate a particular performance, the items should be about evaluating that performance, not actually performing the tasks. It is also important that item styles not be widely mixed in a test, so as to avoid measuring test taking skill instead of subject-matter competence.

Objectivity of Measurement. Third, I would like to mention objectivity of measurement. Each of Frederiksen's categories described above, appears to possess both advantages and disadvantages. Optimally, one would hope to assess individual performance at the highest possible level of fidelity. Unfortunately, this may imply a subjective (rating) technique for a specific situation, which then requires a subjectivity vs. fidelity tradeoff. In order to minimize subjectivity in a real life situation, it may be necessary to decrease the level of fidelity so that more objective measurements (such as time and errors) can be obtained. Such a fidelity decrease can, in certain instances, be theoretically justified. Presumably, an actual increase in overall criterion adequacy may result from a gain in objectivity which compensates for a corresponding loss in fidelity.

In low fidelity performance testing situations, such as those using paper and pencil multiple-choice formats, objectivity in scoring is apparent--such tests can for example, be computer scored. In higher

fidelity testing situations, it is relatively simple to maximize objectivity in so-called "hard-skill" areas such as electronic maintenance. In "soft-skill" areas, such as creativity, leadership, etc. objectivity in scoring is considerably more difficult to achieve. To the extent that objectivity is not achieved, reliability is attenuated.

One suggested method of maximizing objectivity in "soft-skill" testing, is to require several examiners to assess each individual. Inter-rater agreement can then be calculated. If low inter-rater agreement is found consistently, the test should be revised.

Scoring Problems. Fourth, allow me to mention scoring problems in the development of performance tests. The difficulties associated with scoring performance tests, have been described by so many for so long, that by now virtually everyone with an interest in this area knows that problems often include expense, long administration times, apparatus which may break down at inconvenient times, narrow applicability, unreliability, etc. Yet we must develop performance evaluation systems which minimize these difficulties while providing valid measures of performance. Two scoring questions continually arise in performance testing. These concern product vs. process scoring, and the question of assistance vs. non-interference.

Products versus processes. Should "products" or "processes" be scored? Should the extent to which a "right answer" is obtained be measured, or should the extent to which the proper procedure was used be measured, regardless of the final result; or some combination of

these? One way to score within-stage troubleshooting for example, is to determine whether the subject is or is not able to identify a defective component. This method scores only the product of troubleshooting. If such a scoring scheme is used, it is difficult, if not impossible, to determine which of the many possible causes resulted in failure to solve the problem. The subject may have made errors in the use of technical data; he may have made errors in the use of test equipment; or he may have made logical errors in deciding where to make the check. Observation of the performance process may enable identification of the causes for failure.

Another area of concern in scoring products alone, is that there may be only a single task in the task category. If only the product of that task is observed, only a single measure is obtained on each subject for that task category.

Finally, for some tasks, there is no product at the end of the process. Checkout procedures for example, may include energizing the equipment to be checked, making all the required checks, and deenergizing the equipment. If performance of the process is not measured, it is impossible to determine whether the procedure has been done correctly--and this is the primary item of interest.

Three conditions under which processes should be scored in addition to, or instead of, products are: When diagnostic information is required, when additional scores are needed for a particular task, and when there is no product at the end of the process. For an excellent discussion of process versus product scoring, see Osborn (1973).

Assistance versus no assistance. Should an "assist" or "non-interference" method of scoring be used? If the non-interference method of scoring is used, serious distortions of scores may result when inexperienced students are tested. In some cases it may even be impossible to find out how much of the task a person can perform because many of the subtasks require proper performance of previous steps. If the tester does not in some way assist the task performer in step 1, it may, in effect, be impossible to administer the test, even though the examinee may be able to perform all of the remaining steps. The intervention of the administrator does indeed introduce distortion into the meaning of the test score. A slightly distorted score, however, is better than no score at all. If assists can be kept to a minimum, the distortion is likely to be relatively minor. Properly controlled, an assist approach can indeed be used effectively. The nature of many activities is such that an assist method may be mandatory.

Reliability and Validity. Finally, allow me to discuss the areas of reliability and validity in criterion-referenced measurement. Persons who have completed an introductory course in psychometrics understand that the validity of a test cannot exceed its reliability. But to what extent are these traditional concepts applicable to criterion-referenced testing?

Reliability. Stanley (1971) has described techniques for applying traditional reliability concepts as developed in norm-referenced

contexts, to criterion-referenced tests. Since criterion-referenced measures are designed for situations where discriminations among persons are of minimal importance, traditional concepts of test reliability are less applicable. This is the case since criterion-referenced measures are often used in situations having little or no variation among true scores. However, since the basic concept involved is to discriminate individual variation from a fixed criterion score, a criterion-referenced test can give reliable scores even though the classically defined parallel forms reliability coefficient is low.

A recent work by Livingston (1972) has shown how classical concepts of reliability can be applied to criterion-referenced measures. Basically, the procedure involves a redefinition of variance, covariance and correlation in terms of deviation from a criterion, rather than from the mean. Livingston has also shown how other classical norm-referenced reliability concepts, e.g., correction for attenuation and the Spearman-Brown formula, apply to criterion-referenced measurement.

Such techniques are, for the most part however, not fully developed. (For example, see Oakland, 1972; Haladyna, 1974; and Woodson, 1974). The need for additional work in the area of criterion-referenced reliability, continues to be a pressing one.

A practical solution is to assess test-retest reliability of criterion-referenced tests; a procedure which does not depend on internal consistency, and which increases the variability of the test results, because of the two test administrations required. The ϕ coefficient

is useful for analyzing the resulting four-fold (first administration-second administration, vs. pass-fail) data. It has elsewhere been suggested (Swezey and Pearlstein, 1975) that Ø values of less than +.50 tend to indicate unacceptable test-retest reliability for criterion-referenced tests.

Content validity. The process of determining performance criteria on the basis of information obtained directly from job required skills, defines a content-valid criterion. Criterion tests which are derived from appropriate training analyses provide the best available measure of behavioral objectives. No better criterion exists upon which to validate these instruments.

Cronbach (1971) has treated the case of criterion-referenced content validity in his discussion of performance testing. Content validity is a matter of the extent to which a test corresponds to the population performance objectives. Content validation can be viewed as absolute measurement, thus the score on a test suggests that an individual does or does not possess the abilities to adequately perform the task. Cronbach uses the example of a dictated spelling test which, he says, is "a measure of hearing, and spelling vocabulary and ability to write" (1971, p. 453).

Content validity is also temporary. Content valid items reflect behaviors, tasks, etc. which occur in the world today. These change with the passage of time. It is necessary therefore, in developing

objectives-oriented, criterion-referenced tests; that procedures be developed which insure that a prospective user who follows the specified procedure today, will arrive at a test reasonably like the job today. The entire process may change tomorrow.

Content validation, it is argued, is an especially appropriate method in criterion-referenced applications. A test is content valid if the test items are carefully based on the performances, conditions, and standards specified in the objectives; and if the test items appropriately sample objectives. (Of course, the objectives themselves must be sound.) Thus, in most instances, careful test construction will, itself, enable the development of content valid tests. However, in instances where low fidelity tests are constructed, it may be more difficult to determine content validity, since the items are not likely to be precisely matched to objectives. In such cases, there are two additional types of criterion-related validation that are well-suited to criterion-referenced measurement: concurrent validity and predictive validity.

Concurrent validity. In determining concurrent validity, test results are compared with an outside measure of the behaviors tested. This outside measure must be the best available assessment of performance on the objective(s) in question. The assessment of concurrent validity, involves individual assessment via the test and the outside measure close together in time (concurrently). Ø again may be used on the four-fold data (CRT-other measure, vs. pass-fail).

Predictive validity. Performance prediction, using criterion-referenced measures is no less practical or more difficult than is prediction using standard, norm-referenced measurement techniques. Although criterion-referenced scores are often of the "go, no-go" variety, they can be employed as predictors of continuously measured criteria via point biserial and biserial techniques; and of dichotomous standards via phi-coefficients and tetrachloric coefficients. (See McNemar, 1962 for a discussion of these techniques.) Predictive validity is a particularly appropriate concept in the case of criterion-refernced measerement.

Predictive validity involves the same assumptions as does concurrent validity. The outside measure must be an accurate measure of the performance in question, or the validation will be meaningless. Predictive validity can be calculated the same way, except the outside measure is taken at a later time--i.e., when the individuals are actually performing the activity for which they've been trained.

Summary. This paper has attempted to present and discuss some cogent issues in the development of objectives-oriented, criterion-referenced measurement systems. The problems in these areas have not been solved by a long way. Much work remains. Nevertheless, it is suggested that domain-oriented and norm-referenced systems, while appropriate in many situations are inappropriate or insufficient in others. Development of objectives-oriented, criterion-referenced tests must, of necessity, proceed. Guidance in how to construct such tests is continually being developed and distributed. This guidance is based upon the best available experience and the existing state-of-the-art. Yet many fundamental questions remain.

References

Cronbach, L. J. Test validation. In R. L. Thorndike, Educational measurement. Washington, D.C.: American Council on Education, 1971.

Frederiksen, N. Proficiency tests for training evaluation. In R. Glaser (Ed.) Training research and education. Pittsburgh: University of Pittsburgh Press, 1962, 323-346.

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

Glaser, R. and Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement. 1974, 11(2), 93-99.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9(1), 13-26.

McNemar, Q. Psychological statistics. New York: Wiley, 1962.

Oakland, T. An evaluation of available models for estimating the reliability and validity of criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Osborn, W. C. Process versus product measures in performance testing. Paper presented at the Annual Conference of Military Testing Association, San Antonio, October, 1973.

Pearlstein, R. B. and Swezey, R. W. Criterion-referenced measurement in the Army: Development of a research-based, practical test construction manual. Reston, Va.: Applied Science Associates, Inc., November 1974. 308-AR18(2)-FR-1174-RBP.

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

Sanders, J. R. and Murray, S. L. Alternatives for achievement testing. Educational Technology, March 1976, 17-23.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.) Educational Measurement, Washington, D.C.: American Council of Education, 1971.

Swezey, R. W. and Pearlstein, R. B. Developing criterion-referenced tests. Catalog of Selected Documents in Psychology, Spring 1975, 5, 227, Ms. 918.

Swezey, R. W., Pearlstein, R. B., Ton, W. H. and Mirabella, A. Contemporary views on criterion-referenced testing. Reston, Va.: Applied Science Associates, Inc., 1974.

Woodson, M. I. C. E. The issues of item and test variance for criterion-referenced tests. Journal of Educational Measurement. 1974, 11(1), 63-64. (a)