DOCUMENT RESUME

ED 096 335 TM 003 941

AUTHOR Schwartz, Howard P.

TITLE Testing for Instructional Purposes: Norm Referenced--Criterion Referenced.

PUB DATE [74]

NOTE 12p.; Paper presented at the State Convention of the Kansas Personnel and Guidance Association

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE

DESCRIPTORS Comparative Analysis; *Criterion Referenced Tests; Educational Assessment; Educational Needs; Instruction; *Norm Referenced Tests; Test Reliability

ABSTRACT

Distinction between norm referenced and criterion referenced tests are explored in relationship to underlying philosophy and intent. In considering the use of a criterion referenced test for instructional purposes, consideration is given to: specification of objectives, item content and selection, reliability, and needs assessment. (Author)

ED 096335

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TESTING FOR INSTRUCTIONAL PURPOSES:
NORM REFERENCED - CRITERION REFERENCED

Howard P. Schwartz, Ed.D.
Associate Director, Data Processing and
Educational Measurements Center
Emporia Kansas State College

Paper presented to the Kansas Personnel
and Guidance Association State Convention

TM 003 941

Underlying Philosophy: Intent

A reasonable starting point for selecting the most appropriate type of assessment for instructional use is to determine the intent for which the assessment will be used.

At the most basic level the intent of norm referenced measurement is to ascertain how an individual performed in relationship to the performances of other individuals on the same measuring instrument. When there is a need for comparative data, e.g., which students should consider a college preparatory program, or a situation exists where a degree of selectivity is required, e.g., identifying the top 25 students for a new science program, a norm referenced test will best provide this type of data.

The intent of criterion referenced measurement is to provide information which can be related to specific objectives and specific standards of performance. Criterion referenced measurement can be distinguished from norm referenced measurement in that criterion referenced tests do not focus on the problem of individual differences, and are not developed with the intent of determining an individual's relative performance in some reference group.

If the intent of assessment for instructional purposes, is to describe the performance of individuals and groups and to relate that description to judgments of adequacy, standards of performance or mastery levels, criterion referenced assessment would be most appropriate.

This distinction of describing rather than comparing the performance of individuals and groups has important implications for the classroom teacher. If, for example, 8th grade students at a particular junior high school were administered a norm referenced mathematics test, one could refer to the norms table and determine what percentage of students in a defined population scored lower or higher for any given raw score. However, this norm referenced test would not provide information indicating how much each student does know, and one could not make direct inferences about what an individual can or cannot do. While teachers are interested in how well a student does relative to other students, they also recognize the fact that questions need to be answered concerning the number of their students who have learned enough of a subject to have satisfied the minimum objectives of instruction.

Thus, both norm referenced and criterion referenced tests can be used to focus on decisions regarding individuals. However, it is the context within which these decisions are made that really provides the distinction. Criterion referenced measurement, with its intent on describing performance, seems most appropriate for: student evaluation (assessment for the purpose of making decisions about individual student learning) and program evaluation (assessment for the purpose of indicating the proportions of students achieving specified objectives). Norm referenced measurement, with its intent on comparing performance, seems most appropriate for: institutional decisions (assessment for the purpose of making a large number of comparable decisions, e.g., selection, classification, placement, public relations etc.) and individual decisions (typically decisions an individual makes about himself, e.g., vocational choice, educational choice, personal).

Since tests used for instructional purposes are usually conceived in terms of the particular curriculum goals of the school, e.g., the teacher's ability to bring about gains in mathematics (student evaluation) or determining the percentage of some group of students who are able to perform a particular task or who have "mastered" a particular objective (program evaluation) a criterion referenced test would clearly be most appropriate.

Specification of Objectives: Steps

The evaluation procedures that a teacher uses in the classroom should be directed not only towards obtaining evidence on the important objectives of instruction but also toward making clear to students what skills, abilities and knowledge are important in the subject matter area.

The objectives written for an instructional area should be stated clearly and be measurable. Steps to follow in writing curriculum objectives include: (1) determine the criteria for selecting the objectives; (2) determine the goals for the objectives and define them in behavioral terms; (3) determine the heirarchy of objectives; (4) outline the activities to implement the objectives; (5) prepare evaluation instruments to assess the objective; (6) define the competency levels for each objective; (7) field test the items; and (8) readjust objectives or items when necessary.

ITEM CONTENT AND SELECTION: Considerations

Once the objectives for a curricular area have been established, the next undertaking is to construct and or select test items to measure the objectives. This is a difficult procedure because of the vast number of test items that might be constructed for any given objective. If, through a sampling quirk, too many easy items are included, a student's mastery might be over-estimated; if too many difficult items are included, a student's mastery might be under-estimated. For example, consider this objective:

> Compute the correct product of two single digit numbers greater than 0 where the maximum value of this product does not exceed 20.

The specificity of this objective is quite deceptive since there are 29 pairs of numerals that meet this standard and at least 10 different item types that could be used to assess student performance.

An item written for an objective should sample as purely as possible the specific domain of behaviors. This sample of behaviors will not be random, but hopefully, it will be representative of the domain. The most important aspect of the item is whether it is sensitive to instruction.

The number of items to construct for each objective is influenced by several factors. Some of these factors are the amount of testing time available, and the cost of making an interpretation error, such as saying that a student has achieved mastery when he has not. The usual practice is to use about 5 to 10 items per objective. This practice stems more from feasibility constraints than any sound foundation in psychometric theory.

The practice of employing a particular passing score, e.g. 80%, only on the grounds of tradition is difficult to defend. It seems unreasonable to require the same level of proficiency for all domains and all individuals just on the basis of tradition.

If the teacher is going to use "judgment" in setting a passing score, five sources of information can be utilized (Millman, 1972). One procedure is to set the passing score such that a predetermined percent of students pass. This procedure is most applicable when the number of students who should be given some treatment or passing score is fixed and the result of evaluation is to select the most proficient examinees.

Another procedure that can be followed is to inspect the items and make a judgment concerning how important it is that the item be answered correctly. Alternatively, a decision might be made that in order to pass the test a correct answer may be given to all the items in one group, that some fraction of the items in the second group must be answered correctly, and that only a smaller fraction of the remaining items need to be answered in an acceptable way.

A third alternative can be applied by examination of the subject matter and if the knowledge and skills are seen as fundamental or prerequisite to future performance, then a high mastery level can be set. A lower passing score can be tolerated when the material is not seen as completing a necessary link in the development of some complex concept or skill, especially if the concepts will be covered again in the curriculum.

Fourthly, all things being equal, a low passing score can be used when the psychological and financial costs associated with a remedial instructional program are relatively high. There should be fewer failures when the cost of failing is high. These "costs" might include lower motivation and boredom, damage to self-concept, and dollar and time expenses of conducting a remedial instructional program. A higher passing score can be tolerated when these costs are not too great.

A fifth consideration involves the measurement error. Since there is an error introduced in estimating a student's proficiency the passing score could be raised to take into account the expected contribution attributed to random guessing. Also, if the test items are suspected to be unrepresentative, it might be wise to raise or lower the standard an additional amount in order to protect against the misclassification error (student passes when he should fail, examinee fails when he should pass).

Reliability

Since the intent of a criterion-referenced test is to demonstrate mastery or non-mastery of explicit behaviorally stated objectives, the criteria that a student's performance on the test is to be evaluated against would be the behavioral objectives. In a paper discussing reliability problems in a criterion-referenced test Roudabush and Green pointed out:

> The user of a criterion-referenced test wants to know in some absolute sense which criterion behaviors, that is, objectives, of those represented in the test that student has mastered and which he has not mastered, so that the student's further study can be directed towards those objectives of importance to him that he has not yet mastered.

The process of evaluating items as they directly relate to specific objectives and the criterion of what constitutes mastery are crucial problems. Usual item analysis procedures seem inappropriate for the following reasons: (1) Unless the test is to become exceedingly long, few items can be used to evaluate any one specific objective; (2) Scores on a criterion-referenced test may contain no variance for a given population, and yet the test may be a good test. On certain criterion-referenced tests, it is possible that all students completing a particular unit of instruction will pass every item. The possibility that scores on a criterion referenced test may have no variance for some population does cast doubt on the relevance of the concept of reliability as defined in classical test theory; (3) A criterion-referenced test should be sensitive to instruction. Reliability in the traditional sense may be of lesser importance that the appropriateness of the decisions made that affect the treatment of the examinees; and (4) A student's score on a given item may not provide useful information about the optimum performance of the item.

Procedures are still needed to evaluate items in which a criterion score based upon student performance on the items can be determined under these conditions: (1) when the item is administered before instruction on relevant objectives (pre-assessment item characteristics), (2) when the item is administered immediately after the teaching of the instructional objectives, and (3) when the item is re-administered after a passage of a minimum time period (retention characteristics).

Potpourri:

Schools that are considering adopting a system of evaluation that can be directly utilized to facilitate decisions concerning: curricular programs, course refinements, and student achievement may first want to begin implementation by conducting a needs assessment.

A needs assessment is based upon the notion that the relevancy of education must be empirically determined and should identify "what is" and "what should be." There are four basic activities involved in a needs assessment. These activities include:

1. A listing of the full range of possible goals that might be involved in the needs assessment.
2. Determining and ranking the relative importance of the goals.
3. Assessment to determine to what degree the most important goals are being achieved by current school programs (e.g. identifying discrepancies between desired and actual outcomes.)
4. Deciding which of the discrepancies between present and desired performance are the ones most important to correct.

In conducting a needs assessment some basic principals to consider:

1. Involve the local community.
2. Specify and define relevant goals.
3. Instruments used for assessment must have face and content validity.
4. Take the noncognitive effects of school into account.

5. Data presentations should be designed for lay understanding. The wording should not represent a new low in educational jargon,

6. Assessment must not be an end in itself.

Once educational goals have been established through a needs assessment strategy, curricular and instructional objectives can be delineated. When translating curricular objectives into instructional objectives, take these elements into account:

1. Who is supposed to perform the desired behavior?

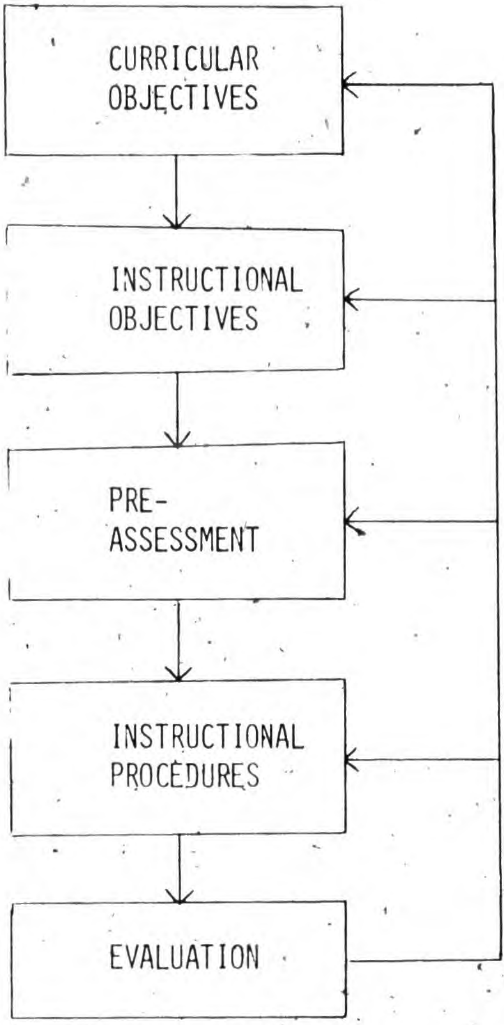
2. What is the actual behavior to be used in demonstrating mastery of an objective?

3. Specify the result of the behavior (e.g., the product or performance) which will be evaluated to determine whether the objective is mastered.

4. Determine the conditions under which the behavior is to be performed (e.g., a 50 minute quiz, oral recitation in class).

5. Set a standard which will be used to evaluate the success of the product or performance (e.g., 80 percent correct, or 7 out of 10).

Taking the broad unmeasurable goals derived through a needs assessment and refining those goals into curricular and instructional objectives, a general model of instruction can then begin to be implemented. Since the goal of instruction is to maximize the efficiency in which all students achieve specified objectives, a general model of instruction that can be utilized in the classroom is offered:


CURRICULAR OBJECTIVES
INSTRUCTIONAL OBJECTIVES
PRE-ASSESSMENT
INSTRUCTIONAL PROCEDURES
EVALUATION

REFERENCES

Baker, E. L. Using measurement to improve instruction. Paper presented at Convention of American Psychological Association, Honolulu, Hawaii, 1972. ED 069 762.

Klein, S. P. Evaluating tests in terms of information they provide. Evaluation Comment, 1970, 2 (2) 1-6. ED 045 699

Millman, J. Passing scores and test length for domain referenced measures. Paper presented at AERA, Chicago, Illinois, 1972. ED 065 555.

Roudabush, G. E., & Green, D. R. Some reliability problems in a criterion-referenced test. ED 050 144.