THE PROBLEMS AND POTENTIALS OF CRITERION-REFERENCED TESTING

By
A. O Afolayan

ABSTRACT

This paper attempts to synthesize in greater detail some of the issues and problems which have been the subject of much discussion and controversy on the development and use of criterion-referenced test. The various steps which have been taken to solve some of the problems in order to maximize the advantages of using the test are discussed. The paper concludes by suggesting that appropriate steps be taken to introduce the use of it for continuous assessment in Nigerian Primary and Secondary Schools.

INTRODUCTION

As instruments for assessing individual differences along a given dimension of behaviour, tests are important in key areas of life. They are most commonly used to evaluate students' achievement and the effectiveness of an educational programme. Beyond the narrow confines of the classroom, they are used to determine personality characteristics such as intelligence, motivation, aptitude, readiness and creativity. Also in a world dominated by work, tests are powerful instruments for administrative decisions concerning selection, appointment, placement, evaluation of job performance and promotion. Other areas of use include counselling and research in and outside the classroom.

The use of tests is so pervasive that they can be said to be overused. They are also often abused and misused. Consequently, they have become much of a bogey to students, a cog in the wheel of much purposeful learning and a source of public discontent against formal education. They are the aspect of the educational system most feared by students and most criticised by the public: yet in the absence of any better alternative, tests and testing cannot be abolished. The shortcomings of test and testing have, however, continued to stimulate efforts in the direction of developing new and better tests and improving testing situations and procedures.

A good example is the development of criterion-referenced test which has been a subject of much discussion and controversy over the last fifteen years. Badmus (1977) has attempted a brief synthesis of some of the issues and problems on which much of the discussion and the controversy have centred. These are issues and problems concerning definition of the term, the differences between the test and the classical norm-referenced test, item construction, analysis and selection, establishment of reliability and validity, and the uses of the test. The purpose of this paper is to synthesize in greater detail these issues and problems as well as the various steps which have been taken to solve the problems in order to maximise the advantages of using the test. The paper concludes by suggesting that appropriate steps be taken to introduce the use of it for continuous assessment in Nigerian primary and secondary schools.

Defining the Term

The term criterion-referenced test (CRT) as it is now commonly referred to and used, was introduced by Glaser (1963). Defining it he said:

*What I shall call criterion-referenced measures depend upon an absolute*
standard of quality... Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others. (1963, pp.519 - 520).

Since the term was introduced, considerable attention has been given to the study of criterion-referenced test in terms of construction, item analysis, reliability and validity. Depending on the individual's creation and/or the aspect of test and measurement intended for emphasis, the term has come to be variously defined.

First, the term has been defined as a sample of population of items. According to Harris and Steward for example.

A pure criterion-referenced test is one consisting of a sample of production task drawn from a well-defined population of performances, a sample that may be used to estimate the proportion at which the students can succeed. (1971, p. 1)

Kriewall's (1970. p. 5433A) definition of the term as "a sequence of Bernoulli trails on items randomly selected from a well defined population of items" is another good example. The problem with this definition as Loevinger (1965) has pointed out is that one cannot define a population of items and sample randomly from them except in some unusually restrictive situation. For this reason, this definition has been described as representing an extreme position: or as Hambleton and Novick (1973) have put it, the most restrictive definition that has been proposed

Secondly, the term has been defined in terms of difficulty index. According to Dziuban and Vickery (1973, p.483), Kriewall has defined the test in this way as "one in which items are homogeneous in difficulty for each examinee". Emrick (1971) even goes further to define it as a group of items characterised not only by equivalent difficulty, but also equality in form and content. The problem here is with the term "difficulty index". As would be seen later, it is an item analysis technique for the traditional norm-referenced test; and, therefore, cannot be used except with great caution in defining criterion-referenced tests.

Thirdly, the term has been defined in relation to behavioural objectives. According to Iven (1971, p. 4548A) for example, a criterion-referenced test is a test "composed of Items keyed to a set of behavioural objectives". This definition does not really tell much about the test itself, nor about behavioural objectives which it is said to be tied to. It is therefore. Probably the least Informative of all the definitions that have been proposed.

Lastly, criterion-referenced tests have been defined in terms of performance standard. Millman's (1970, P.226) definition of the term as "those measures which are used to ascertain an individual's status with respect to some criterion, i.e., performance standard" is a very good example of this type of definition, so also is Glaser's definition earlier referred to. Others with similar definitions include Glaser and Nitko (1971), and Livingston (1972). These definitions not only emphasize the use or the purpose of the test, they also suggest the process of developing it and interpreting whatever scores students make on it. It seems, therefore, unnecessary to make any distinction between Livingston's definition and that of Glaser and Nitko as Badmus (1977) has done. As these definitions seem to point to the major differences between the test and the classical norm-referenced test, they are generally regarded as very flexible and useful definitions.

Differences between Criterion-Referenced Tests (CRT) and Norm-Referenced Tests (NRT)

The various definitions that have been examined will, in no small measure, help to illuminate much of the controversy that has developed around criterion-referenced tests. It is however by examining some of the basic differences between criterion-referenced tests (CRT)
and norm-referenced tests (NRT) that the issues and problems can be more fully comprehended. One of such basic differences concerns the primary purpose which each of the two tests is meant to serve.

There seems to be a consensus that one cannot distinguish a criterion-referenced test from a norm-referenced test simply by looking at them. It is generally agreed, however, that the two types of test are meant to serve different purposes. Cox and Graham (1966), Millman (1970), and Gorth and Hambleton (1972) have all emphasised that while the primary purpose of a norm-referenced test is to compare individuals, the main purpose of a criterion-referenced test is to make decisions in terms of the individual student's performance with respect to some criterion. Popham and Husek (1969) have also maintained that norm-referenced tests are used where a degree of selectivity is required, while criterion-referenced tests are used where one is interested in whether an individual possesses a particular skill with no selectivity involved or constraints put on the number that can possess it. This difference has been interpreted in various ways. It has, for example, been interpreted to mean that criterion-referenced tests can be used to serve a more diagnostic purpose than norm-referenced tests. It has also been interpreted to mean that the A, B, C grading system cannot be used referenced tests because they are based on ranking, that is, on comparison.

Another point of difference which cannot be said to belong in the group of further distinctions which Carver (1974) and Badmus (1977) regard as misleading is the process by which each type of test is developed. The process in norm-referenced tests is usually that of standardisation on a sample otherwise known as the normative group. Individuals that take the test afterwards are then compared with the normative group, hence the term "norm-referenced test." In constructing or developing criterion-referenced tests mere is no such standardisation. Rather, each test is developed in such a way that individuals taking the test can have further instruction after each unsuccessful attempt and then take the test again until they reach a pre-specified criterion or performance level.

The process of standardising a norm-referenced test on a normative group involves "spreading out" the sample used for the standardisation so that people taking the test can in, like manner, be spread out. It is only when the test helps to spread people out for comparison that the test can be said to meet the primary objective for which it is designed. In other words, the spreading out of people which, in statistical language, is usually referred to as variance is a necessary condition for norm-referenced tests. As the purpose of criterion-referenced tests is not to compare individuals, the test needs not spread people out. An important distinction between the two tests, therefore, is variance. In fact, as Popham and Husek (1969) have suggested, this distinction may very well be the core of the differences between the two tests. It must be emphasised that one can and, in most cases, will obtain variance from criterion-referenced tests as one will obtain from most norm-referenced tests. The point, however, is that variance on a criterion-referenced test is not the basis for interpreting scores, because the question is whether each student has reached the performance standard or not. In other words, if variance is obtained at all from any criterion-referenced test, the variance cannot be explained and/or applied from the point of view of the classical norm-referenced test theory.

Item Analysis

No test can achieve the purpose for which it is intended unless its items go through a process of careful selection. This process particularly in the traditional norm-referenced test often involves complex statistical analysis of items that have passed through some kind of trial testing. As norm-referenced and criterion-referenced tests are intended to serve different purposes, using the traditional norm-referenced item analysis and selection techniques for criterion-referenced test is most likely to be conceptually and statistically inappropriate, some of the problems that are
likely to arise from using such traditional techniques for criterion-referenced tests are worth examining.

One of the most commonly used techniques of item analysis in norm-referenced tests is difficulty index. Describing the term as a misnomer, Nunnally (1972, pp 186 — 187) uses instead the term "easiness percentage" and explains it simply as "the number of students who get an item correct divided by the total number of students taking the test". He asserts that only items with fifty percent easiness percentages can provide maximum discriminations, but suggests that items with easiness percentages between twenty and eighty can be selected.

Since the primary purpose of using criterion-referenced tests is not to discriminate among students, it is obvious that difficulty index or easiness percentage cannot be used with criterion-referenced tests in the same way as it is used with norm-referenced tests. In most cases, criterion-referenced tests are constructed so that each item can measure the accomplishment of a particular objective. It follows, therefore, that if objectives are well written, instructional strategies are appropriate, and the list items are carefully constructed to measure the objectives, items measuring the same objectives must have about the same difficulty index. Otherwise, either the objectives, the instructional strategies, the test items or any combination of the three needs re-examination. Items with difficulty indices of ninety to one hundred per cent will, for not discriminating among students qualify for automatic rejection in norm-referenced tests. They will be the ideal items for criterion-referenced tests, because a vast majority of the students for which the items were written will be able to pass them and, therefore, achieve the objectives on which the items have been written.

Another item-analysis technique that has been traditionally used with norm-referenced tests is discrimination index. First, the number of students in the upper and lower twenty-seven per cent according to Cox (1965), or to Cox (1965), or twenty-five per cent as suggested by Nunnally (1972), is determined. Then, the percentage of students in each group who pass each item is calculated and the difference between the two percentages for each item is found. In the case of norm-referenced tests for which this technique is traditionally used, the good items are those with high positive difference. On the other hand, items with zero difference will probably be the most acceptable in criterion-referenced tests. It appears, therefore, that unless the technique is modified, it cannot be applied with any degree of confidence to criterion-referenced tests.

The limitations in applying the traditional test-analysis technique to criterion-reference tests have stimulated some interest in the development of procedures that can be more appropriately used with the test. In trying to develop one, Cox and vargas, (1966) compared the traditional discrimination index with pre- and post test discrimination index. The pre — and posttest discrimination index was computed by subtracting the percentage of passes on an item on the pretest from the percentage of passes on the item on the posttest. They concluded that:

"The pre — and posttest method of item analysis produced results sufficiently different from the traditional method to warrant its consideration in those cases where score variability is not the concern, such as in criterion-referenced measures" (Cox, 1971, pp 71 — 72).

Gorth and Hambleton (1972) have tried to approach the problem about the same way as Cox and Vargas (1966). Instead of comparing the traditional discrimination index with the pre- and posttest discrimination index, they computed a pre- and posttest bi serial correlation for each item, and calculated a pre- and posttest discrimination index, and pre- and delayed posttest discrimination index for each item. Three correlation coefficients were then computed from the three sets of data. From the coefficients obtained, They found that the correlation between the pre-and posttest on the one hand and the pre-and delayed posttest on the other is higher than the other two. The conclusion of Cox and Vargas (1966) that the pre- and the posttest discrimination index can be effectively used to select items for criterion-referenced tests was thus supported.

As in the case of the discrimination index, some attempts have been made to modify the difficulty
index to suit criterion-referenced tests. Iven (1971) for example, computed pretest-post test difficulty values, that is, the difference between the number of correct responses for each item on the pretest and the posttest, and found the procedure appropriate in selecting items for criterion-referenced tests. He, therefore, concluded that, "item selection for criterion referenced tests could be accomplished most efficiently using the difference in pretest-posttest difficulty values." (1971 p. 4548A).

It is important to note that the different modifications of the traditional techniques that have been attempted for selecting items for criterion-referenced tests are applicable only to pre- and posttest conditions. They are not applicable where only one test administration is possible. Item analysis when only one test administration is possible, therefore, remains a problem in criterion-referenced testing.

Reliability

Reliability refers to the consistency with which a test measures what it is intended to measure. The traditional techniques for estimating whether a test meets this requirement include test-retest, equivalent or parallel form, split-half and Kuder-Richardson Formula 20. These traditional techniques are based on variance. Consequently, they are inappropriate for estimating the reliability of criterion-referenced tests for which variance is not a necessary condition.

As in item analysis, the inappropriateness of the traditional techniques has stimulated the trying of new ideas and techniques. One such new idea that has been suggested and tried is replicability. Carver (1970) suggested the idea by pointing out that reliability of any test depends on replicability which itself is not dependent on test-score variance. He maintained that similar scores on parallel forms of a test is near perfect replicability and, therefore, means reliability for criterion-referenced tests, even though the reliability of the same test by any of the traditional methods would be close to zero. The idea was put to test by Iven (1971). Using items which about the same number of students got right on a pretest and a posttest, he developed a test and constructed a parallel form of it. He administered both three times as pretests, posttests and re-tests, and found that the scores remained similar across posttest - re-test and post test-parallel form administrations, thereby showing that replicability can be used to establish reliability for criterion-referenced tests. One problem, of course, as in the case of item analysis, is that the technique cannot be used when only one test administration is possible.

An attempt has been made by Cox and Graham (1966) to solve the reliability problem by applying the Guttman's scalogram technique. A sequential criterion-referenced test was constructed to measure a set of objectives that was equally sequential. As expected in this type of test, each student was able to answer all items up to his level of ability correctly and miss all items beyond that point. After the test had been revised and Menzel's procedure applied to improve its scalability, a reproducibility index of .970 and a scalability coefficient of .792 were obtained. As the reproducibility index is an indication of how well an individual's response pattern can be reproduced from a knowledge of his total score, it was suggested that the index can be used as a reliability estimate for criterion-referenced tests that are sequential. Tests of this type are, of course, very rarely constructed. The technique is therefore, likely to be of very limited use.

A different approach based on the assumptions of the classical test theory, particularly the assumption of unidimensionality has been suggested by Livingston (1972). He substituted variance from the criterion score for variance from the mean score, and mean products of deviations from the criterion score for co-variance. He then proceeded to derive algebraically what he called criterion-referenced reliability. As the approach is based on the assumptions of the classical test theory, it has not been favourably received. Harris has argued that "the coefficient is based on two populations with means equally distant above and below the criterion score." (1972), p.27). He therefore, concluded that Livingston has not advanced reliability theory for the
special case of criterion-referenced testing. Shavelson Block and Ravitch (1972) have also argued that Livingston's coefficient is unnecessary because conventional reliability statistics are appropriate for his tests since they satisfy the assumptions of the classical test theory.

Hambleton and Novick (1973) have not only rejected Livingston's approach, they have also proposed an alternative based on decision theory. The approach requires estimating the threshold loss, that is, estimating each subject's true score in relation to the criterion or the performance standard. In this way, the actual number of students that reach the criterion level on test administered pre and post or on parallel forms can be determined and compared. Though the technique does not appear to suffer any of the limitations of the classical test theory, its suitability remains to be adequately demonstrated.

Validity

Another area of importance in test development is validity by which we mean the extent to which a test measures what it purports to measure. It is essentially what determines the usefulness of a test and, therefore, is equally important, if not more important than reliability. Yet much less work has been done on the validity of criterion-referenced tests. However, as Popham and Husek (1969) have emphasised, a criterion referenced test must above all else have content validity. This can be accomplished by having a table of specifications for the content area; and the skills, abilities and thinking processes to be tested stated in form of behavioural objectives. The test can then be constructed such that the items form a good representation of the specifications in the table.

Attempts have been made to establish for criterion-referenced test two other types of validity not mentioned by Badtui (1977). The first is predictive validity. Hambleton and Novick (1973) have suggested that this type of validity can be established for the test using the same technique they have proposed for reliability. They maintain that decisions made on one test through the use of the threshold loss can serve as predictors of performance on a new test which need not have the same performance standard as the first test. However, as earlier pointed out, the technique has to be tried much more to establish its suitability. The second is construct validity which has been suggested by Popham and Husek (1969). Their concept of construct validity for criterion-referenced test is, however, that of the traditional concept of intercorrelation among several predictors of the same criterion. Hence, their suggestion may be fraught with serious limitations.

Usefulness

The introduction of criterion-referenced tests has been largely a response to the demand for tests that are appropriate for individualised instructional programmes designed to help students learn at their own rate. Programmed instructions such as Computer Assisted Instruction (CAI) and Individually Prescribed Instruction (IPI) are good examples of such individualised instructional programmes. Their common features include specification of instructional intent in form of behavioural objectives, careful diagnosis to determine each student's point of entry to instruction, provision of multiple instructional modes, individual pacing and sequencing of materials, continuous evaluation to determine the attainment of objectives, and careful monitoring of each student's progress.

The relationship of criterion-referenced tests to such individualised instructional programmes which emphasise in particular the use of behavioural objective has led to some controversy over the usefulness of the test.

Some authorities like Gorth and Hambleton (1972) have argued that criterion-referenced tests are useful because of their advantages in terms of individual assessment and programme
evaluation. Brazziel (W2) elaborates on their usefulness by discussing their specific advantages. He maintains that criterion-referenced tests have some advantages:

1. Permit direct interpretation of progress in terms of specified behavioural objectives.
2. Facilitate individualised instruction.
3. Eliminate a situation where half of school children must always be below the median.
4. Enable the teacher to check on the student's progress at regular intervals.
5. Eliminate pressures on the teacher to "teach to the test."
6. Enable teachers to compile a comprehensive record of each child's development.

Others maintain that though criterion-referenced tests have some advantages, they cannot make any real difference in teaching and learning, because of certain limitations. Ebel (1971), for example, concedes that criterion-referenced tests have some value which norm-referenced tests lack but insists that they cannot significantly improve evaluation of educational achievement because:

1. They do not tell us all we need to know about achievement, for example, excellence and deficiency in learning which can only be measured normatively.
2. They are difficult to obtain as they require detailed specification of objectives or outcomes in behavioural terms.
3. They are necessary for only a small fraction of important educational achievements as they are inappropriate for measuring diverse talents, interests, abilities, understandings and appreciations that keep a society functioning and are possessed in varying degree by all.

These limitations do reflect a bias against behavioural objectives to which criterion-referenced tests are more or less tied. In other words, they seem to be indirectly concerned with criterion-referenced tests. It is in this vein that the limitations have been critically examined by Block (1971) and dismissed as reflecting an unjustified pessimism. This is not to say that criterion-referenced tests suffer no limitations. There are as has been shown, measurement and/or statistical problems concerned with item analysis and selection, reliability and validity. There is even the problem of how to develop test items that can measure complex learning particularly in the affective domain. There is also the problem of the choice of the criterion or the performance standard which is often made arbitrary. These are real problems, but most if not all of them can be solved as teachers gain experience in the development and use of the test. A few of the problems may prove intractable, but the advantages of using the test to help each student grow, learn and experience success at his own rate far outweigh such problems.

Conclusion

Perhaps at no other time in the history of Nigerian education is there a greater need to improve our standard of education, especially at the primary and secondary school levels. In the last five years, primary and secondary school populations have increased so much that many capable primary school pupils cannot be placed in secondary schools because of lack of accommodation. Over the same period, we have witnessed a very high rate of failure, particularly at the secondary school level where the rate has consistently been over 50 per cent. The feeling in many circles is that such a high rate of failure is an indication that our system of education is long overdue for real transformation.

The overall level of students performance at any level of the educational system depends largely on the quality of classroom instruction and the methods of assessment and evaluation. It is, therefore, most probably in these areas that the most urgent measures should be taken to improve our standard of education. The new national policy on education seems very clear as to what these urgent measures are. The most important of them perhaps is to introduce and
positively encourage new methods of instruction that will help the classroom teacher focus or centre instructional activities on the learner. To complement the introduction of new instructional techniques, the present methods of assessment and evaluation by the norm, and the overemphasis on the final examination will have to be jettisoned for a more liberalised system based wholly or in part on continuous assessment of students' individual progress.

As has been pointed out, criterion-referenced tests are particularly used, not only in promoting individualised instructional techniques, but also in assessing the individual progress of students. It seems, therefore, that through the development and use of this type of test, some of the improvements that are most urgently needed to upgrade our standard of education at the primary and secondary school, levels can be effected. In this respect, the teachers' colleges, the National Teachers' Institute, and the institutes and faculties of education have a great role to play, for it is by providing preservice and inservice experiences in these new instructional and evaluation techniques that they can be introduced into our schools.

REFERENCES


