Criterion - Referenced Test Construction and Evaluation

By

Dr. Fidel E.O. Iweka
Department of Educational Psychology, Guidance and Counseling
University of Port Harcourt, Rivers State.

Abstract
This paper is on criterion Referenced test construction and evaluation. Two forms of a 25-item multiple-choice-criterion referenced vocabulary test were developed and administered to two groups of community secondary school Obigwe in Rivers State of Nigeria (n=87) for diagnostic and achievement purposes in a counter balanced pretest/post test design. The dependability indexes for these tests were low or moderate and the item analysis of the criterion-referenced tests suggests there was a slight increase in score gain after a period of 13 weeks of instruction. This suggests that most of the students mastered a modest amount of target vocabulary.

Introduction
Paul (1995) proposed that criterion-referenced tests are procedures device to examine a particular form of communicative behaviour. Criterion-referenced tests do not reference to other children’s achievement but only determine if the child can attain a certain level of performance. Montgomery and Connolly (2002) stated that criterion-referenced tests document individual performance in relation to a domain of information or specific set of skills. Therefore, criterion referenced tests are designed to measure changes in successive performance in an individual. Criterion-referenced tests are used specifically for program planning and evaluating; however, they can also be standardized. Onunkwo (2002) states that criterion-referenced tests are used to determine the degree of proficiency student have attained on some specified objectives or skills. Hanbleton cited in Onunkwo (2002) described it as one which yields data that are directly interpretable in terms of specified performance standard or criterion. Thus, in criterion-referenced measurement or evaluation, a student’s performance is determined by comparing it with a cut-off score or an anchor point earlier set. This type of test is based on absolute rather than relative standards. Its primary use is to measure mastery of specific skills and test items, based on known performance objectives associated with the tests of interest. Criterion-referenced tests are sensitive to and can be used to measure the effects of instruction, based on task analysis, related directly to instructional objectives. Sensitivity is defined as the accuracy with which the test identifies children with problems. (Marrell 2000). The ability to tie the test directly to
the program objectives is another benefit of criterion-referenced tests. Freeman and Miller(2001) reported that criterion-referenced tests were consistently rated as the most useful assessment tool both for understanding the child’s abilities and needs, and for planning teaching responses to them. This assessment tool refers directly to the curriculum, and is likely to be considered comprehensible and relevant.

**Difference between Norm-Referenced and Criterion-Referenced Tests.**

Mc-Cauley (1996) summaries the differences between norm-referenced and criterion referenced tests in a simplistic way. The first difference is the fundamental purpose of both tests. The fundamental purpose of norm-referenced tests is to rank individuals whereas the fundamental purpose of criterion-referenced tests is to distinguish specific levels of performance. A second difference is the test planning. Norm-referenced tests address a broad content and criterion-referenced tests address a clearly specific domain. Lastly, the difference is on how the individual’s performance is summarized. With norm-referenced tests, performance is summarized meaningfully by using percentile ranks and standard scores, and criterion-referenced test performance is summarized meaningfully by using raw scores.

**Similarities between Norm-Referenced and Criterion-Referenced Tests**

Even though norm-reference are criterion-referenced tests have may differences, there are a few similaries for example, criterion-referenced and norm-referenced tests should demonstrate the same interrater and test-retest reliability (Montgomery & Connolly, 1987). Issues of validity, such as content, concurrence, and predictive value should also be similar between the two tests when administered.

Studies comparing Norm-referenced and criterion-referenced tests. Handicapped versus non handicapped: Mactruk and Neiswarth (2000) studied the degree of correspondence between a presumably appropriate norm-referenced device and a project specific, criterion-referenced measure of child progress, as well as the utility of the two different measures in program planning, delivery and modification. The 17 subjects in their study were divided into two groups, handicapped (7) and non-handicapped (10). Handicapped was defined as displaying significant but not profound dysfunctions in at least two of four bases of developmental areas: communication, intellectual, social, and motor. Subjects ranged in age from 19 to 51 months. Inclusion criteria for these children were (a) be ambulatory, (b) be non-handicapped sensorially, and (c) manifest no extreme disorders requiring close clinical or medical supervision.

Both groups of children were administered a norm-referenced and a criterion-referenced test on a quarterly basis. Results of the norm-referenced test were converted into a developmental quotient, a developmental rate, and a mean developmental quotient. Additionally, the criterion referenced tests curriculum objectives were converted to a percentage of total objective achieved.

Through the distribution of scores the researchers found no significant relationship (r=.08 to .43, P< .01) between the total sample of subjects and non-handicapped group with a Spearman rank correlation coefficient. However, the handicapped group demonstrated a high correlation ( r=.93 to .99. p<.01) between the
development quotient and the criterion measures. The Wilcoxon matched pair signed-ranks test was performed to test the feasibility of using the mean developmental quotients as a measure of children progress. This analysis indicated there were no significant differences between the pre and post development quotient for both non-handicapped \((T=25.5, P>.05)\), and handicapped children \((114.0, p>.05)\).

Overall, the findings in this study demonstrated a high correlation \((.93\text{ to } .99, P<.10)\) between the norm-referenced and criterion-referenced tests for the handicapped children. The norm-referenced test was highly correlated \((.90\text{ to } .96, P<.01)\) with classroom measures and had utility as a diagnostic device for educational programming. This high correlation allows the norm referenced assessment results to be a useful measure of a child’s progress once in an educational programme, as the concern of insensitivity is not warranted. These findings supported the following Conclusions:

(a) Norm-referenced developmental quotient and criterion-referenced measures yielded, approximately similar information for handicapped children

(b) Classroom activities showed differential effects for each sample, and

(c) Mainstreaming did not have any apparent adverse effects on the normal children in the sample.

**Purpose of Study**

The purpose of this study is to conduct on intervention construct validity using a pretest/post test design with two forms counterbalanced. Thus, two CRT forms were developed and administered to two groups of students both as pretests and posttests in a proficiency based curriculum.

**Research Questions:**

To guide and direct this study, the following research questions were formulated:

1. To what extent were the two CRT farms dependable in both administrations ?.
2. To what extent did the students master the vocabulary items on the two farms of the CRTS ?

**Method**

**Participants**

This study involved 87 S.S. 3 students of community secondary school Obigwe. \((N=87)\). They took a general English Language class that focused on reading and listening skills. A placement test was administered to make decisions about a proficiency based curriculum in the programme, streaming the students into two course levels. The students were divided into two groups to carry out a counterbalanced design. Since the test involved a listening component and microphones were not used, it was important that students in the same room took the same form of the test. The lower proficiency group was designated group A \((n=44)\) and the midrange group was designated group B \((n=37)\).

**Materials**
Two teachers set the class term objectives by referring to the class goals that had already been set by the administrators of the school. One part of these goals included learning more academic English vocabulary.

The two teachers worked together to design the lesson plans. Each test form consisted of 25 multiple choice items. The target skills for this class were two receptive skills, so the researcher felt that multiple choice items were suitable to test students receptive skills. Six items were included in both forms to help “anchor” the scores. Students were instructed to select the best option (A, B, C or D) which was the closest to the meaning of the underlined target word.

**Test Procedure**

Form A and form B were administered to Group A and Group B as a pretest. The teachers informed the students that their pretest scores would have no effect on their final grades and explained the purpose of the diagnostic test administration. Although test score sheets were returned to students, the question sheets were all collected to avoid information leakage. With respect to the classroom instruction concerning the target vocabulary items, the teachers used the same lesson plans and provided the students with their corresponding meanings. The students were asked to study the vocabulary items included in the lists provided in the class because the words would be tested on the day of the final examination. At the end of the term the test form, that the students did not take at the beginning of the class was administered. Their test scores were used to decide 15% of their final grades. Students were given 15 minutes to complete each test.

**Analyses**

The response of the students were dichotomously scored (converted to correct or incorrect responses) and processed in spread sheets. Missing blanks were treated as incorrect responses. Descriptive statistics for all the items were calculated.

A norm-referenced reliability statistic known as the KR-20 (Brown, 2005) was first employed.

Norm-referenced reliability was used to estimate how much error contributed to the examinees scores. Brown (1990) developed a short-cut formula to estimate the index of dependability, which was also used to estimate the test consistency of CRTS.

Dependability is different from reliability in that it concerns the consistency of absolute decisions, not relative decisions. The coefficient obtained from Brown’s short-cut formula was exactly equivalent to the generalizability coefficient for absolute decisions obtained from a decision study in generalization theory. Two criterion-referenced item statistics were considered especially important: the ID and B-index.

**Result**

Table 1: Descriptive statistics for two farms of a 25-item English vocabulary test administered to two groups of S.S. 3 students in C.S.S Obigwe in 2009.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Referenced Test Construction and Evaluation - Dr. Fidel E.O. Iweka</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>min</td>
</tr>
</tbody>
</table>
Table 1 displays the descriptive statistics. Because Group A (n=44) was less proficient, their mean was 10.14 for the 25-item test; however, Group B (n=37) was more proficient and obtained a slightly higher mean of 13.19—just over half of the test items. The KR-20, a norm-referenced reliability coefficient, respectively yielded 0.66 and 0.40 on form A and form B for the pretests. The dependability indexes were estimated based on the coefficients derived from the KR 20. These indexes of 0.05 and 0.37 were lower than the KR-20 coefficients.

At the end of the term Group B, in which 36 students took form A as posttest, obtained a mean of 17.75 for the test average 71% of the items correct. One student in that group obtained a perfect score of 25. Group A took form B as a posttest and obtained a mean of 12.48—averaging just under half of the items correct. The standard deviation for this group had the largest value of 3.47, indicating that some students studied for this test but others did not. The KR-20 coefficients were 0.49 and 0.57 and the dependabili- ty indexes were 0.46 and 0.53. The data obtained from form A and form B used as pretests and as post tests were added accordingly, and the combined means for the pretests and posttests were 11.53 and 14.85. Thus, a slight increase in the mean scores was observed. With the exception of one Kurtosis value of -1.04, normality was not a problem.

Table 2: Criterion —referenced item statistics for two forms of a 25-item English vocabulary test administered to two groups of S.S. 3 student of C.S.S Obigwe in 2009.
Table 2 summarizes the criterion-referenced item statistics for both forms. Ideally, while IF should be close to 0.00 in pretests, IF should be close to 1.00 in posttests so that DI values can be maximized. The IF values for items 10 and 16 in form A and items 5, 8, 12, 16, 18 and 24, in form B used as pretests, was excessively high, indicating that the students knew the vocabulary items before instruction, the IF values for items 21 and 22 in form A and for items 4, 10, 11, 13, 14, 15, 20 and 21 in posttest form B, were unreasonably low, indicating that the students did not learn the vocabulary items even after instruction. The B-index was calculated by subtracting IF for the bottom 70% of the students from the top 30% of the students. The B-index is sensitive to the location of the test cut-off point. The cut-off point of this test was arbitrarily set as 70%. In the words, student who scored higher than 15 out of 25 points passed the posttest. The B-index values close to 1 indicate an item which has highly differential power, while the B-index values close to -1 signify the opposite. Note that the B-index for all the items in form A taken by group A as a pretest were negative because none of the students exceeded the cut off point of 15. In the other administrations only a few items had negative B-index values. In this study DI was considered the most important criterion-referenced item statistic because it was an indicator of the number of items that the students had learned since the pretest, presumably as a result of instruction. Ideally, DI should be close to 1 showing that the students learned the item.

However, it should be noticed that 14 items in form A had negative DI values. These unwanted results were probably due to the difference in the proficiency levels between group A and group B.

Discussions of Finding
Now, let us reflect on the research questions in terms of the study results.

1. To what extent were the two CRT forms dependable in the two administrations. Based on the phi dependability indexes, with the exception of the form taken by group A as a pretest, the other test forms were found to have moderate dependability values from 37 to .53. Since most of the students in group A scored low on the pretest it was ideal as a diagnostic test since it revealed most students had not yet learned the items. However, because there was not much variance observed in the test scores, the dependability of the test was probably low. Statistics can be an indicator for deciding the quality of items, never the
less, especially when the sample size and criterion-referenced item number are both small, teachers should examine the content carefully in order to decide whether or not items are really measuring the target objectives of the class.

2. To what extent did the students master the vocabulary items on the two forms of the CRT. The pretest/posttest design with two counter balanced forms enable teachers to determine to some degree the effectiveness of their instruction. Such designs focus on two indicators: DI and score gain. To calculate DI, the same items have to be administered as pretests and posttests. Recall that DI for the items in the posttest given to group B had negative values. Because the proficiency level of Group A and group B differed, this was not surprising. Ideally the DI statistic should be used when the proficiency levels of two groups are almost equal. To resolve the problem in this study, each class should have been split into halves.

The other indicator of student learning is score gain. This is a simple but useful method for getting some picture of the effectiveness of curriculum. Recall that the means for the combined pretests and posttests were 11.53 and 14.85, respectively. This suggests that some degree of learning may have occurred in the interval between the tests. That also can be one of the arguments for the construct validity of the CRTs.

Conclusion

The phi dependability indexes of the CRTs were reported and one validity argument was reported. For classroom tests, teachers often use a variety of test formats, but this test relied solely on multiple-choice items. It would be interesting to apply generalizability theory to investigate how much test formats as a facet, can contribute, to the total variance. Since teachers use CRTs frequently in class more studies on CRTs need to be done in the future.

References:


