NORM-REFRENCD TESTS VS CRITERION-REFRENCED TESTS?

Asst.Prof.ABDUL-JABBAR ALI DARWESH
Asst. Prof.EMAN FATHI YAHYA (M Lit.)
UNIVERSITY OF AL-MUSTANSIRIYAH
COLLEGE OF BASIC EDUCATION

Introduction

1.1 The Problem

Since the late 1970s onwards, there has been a general shift of emphasis in approaches to second and foreign language teaching with language being viewed as a tool to be used in purposeful manner. With this shift, new ideas about language testing and new methods of evaluating the performance of classroom learners have emerged, (Weir, 1988). This is so because testing is highly influenced by the methods of teaching. Educationalists and test experts have directed their full attention toward developing tests that have the ability to assess the performance of the learners inside the classroom. Hot arguments have been aroused to decide the appropriate techniques that can be used by teachers to evaluate the achievement of their pupils inside the classroom without any problems. Many ad hoc meetings and symposiums have been held and many committees and boards of examinations have been established to carry out

research and devise competent techniques that assess the performance of learners and pass judgment on broad and narrow curricula at school and elsewhere.

Unfortunately, in Iraq, there are directorates of examinations, but their major task is to keep records of students' results of school examinations and issue certificates for their applicants. They also supervise public examinations. These tasks are very important and even a must; yet, they should not be the only concern of these institutions. In developed countries, boards and directorates of examinations issue many decisions on the methods of assessments and help in the development of real curricula through the feedback they receive as a result of the analysis of the results of the various tests and examinations conducted at different educational institutions.

Since the role of the directorates of examinations has not been influential in deciding on the appropriate methods of assessment, teachers of English in Iraq have been left with no real guide to assess the performance of their students. They tend to mimic public examinations procedures which use more complex techniques that require a number of test experts to design. This is so because public examinations are usually standardized, (Sanders and Horn, 1994) which makes them unfit to be used right away by classroom teachers (Harris, 1967:2).

To adopt techniques that can be prepared and used by the teacher inside his classroom directly, testing specialists have different points of view. Some back the use of norm-referenced tests (NRTs) and find them quite convenient since, 'the quintessential NRT is the standardized test that has tried out with large groups of individuals whose scores provide 'norms' or reference points for interpreting scores' (Bachman, 1990: 7-8), and since the teacher has to compare his student’s performance to that of his class mates or that of another group. Others criticise such tests and prefer using criterion-referenced tests (CRTs) since all tests and quizzes written by school teachers

should be criterion-referenced, (Glaser, 1963), and since CRTs provide information about the degree of mastery of a given criterion domain or ability level of an individual learner,(Bachman, ibid.). The objective is simply to see whether the student has learned the material he is exposed to or not. This is, of course, the ultimate end of any successful teaching process.

The problem of the present study stems from the above-mentioned clashing ideas of testing experts, i.e., whether to use CRTs or NRTs and also from what Oller (1979: xviii) states that most of the textbooks, journals, pamphlets and magazines about classroom language testing techniques are:
intended primarily for learners of FL or EFL, and yet they are generally based on techniques of testing that were not developed for classroom purposes but for institutional standardized testing.

1.2 The Aims

In order to present a solution for the problem which is proposed in the statement of the problem above, the present study aims at:

1. setting a theoretical background that tackles the major characteristics of both NRTs and CRTs; and 2.
drawing a comparison between the two types of tests.

1.3The Hypothesis

To fulfil the above-mentioned aims, it is hypothesized that the merits of CRTs outbid those of the NRTs.

1.4The Delimitations

The scope of this study is delimited to the following:
1.a general discussion of NRTs and CRTs; and
2. a comparison between the two types of tests based on what Popham (1975) has proposed.

1.5The Value

The value of this study springs from the fact that using the proper testing techniques is an end by itself. It provides important information concerning the two major approaches to language testing, i.e., NRTs and CRTs for teachers of English as a foreign language, supervisors and all the people who are concerned with the process of developing language learning and teaching in Iraq.

1.6 The Procedures

To fulfill the aims and verify the hypothesis of the present study, the following procedure are adopted:

1. Conducting a thorough survey for NRTs and CRTs.
2. Drawing a comparison between the two types of testing based on four major aspects, i.e., purpose, content, item characteristics and score interpretation in addition to a fifth aspect, reliability-validity tension.

2.1 Norm-Referenced Tests VS Criterion-Referenced Tests

To meet the criteria of the first aim, this chapter is going to discuss the following topics:

- Norm-Referenced Testing
- Criterion-Referenced Testing
- Reliability and validity of norm-referenced and criterion-referenced testing.

2.1.1 Norm-Referenced Tests

To start with, it might be useful to give a definition of NRTs from (Wikipedia, 2008) which states that:

A norm-referenced test is a type of test, assessment, or evaluation in which the tested individual is compared to a sample of his or her peers (referred to as a 'normative sample').

'Normative' interpretation of a NRT refers to the process of comparing the performance of a tested individual to the performance of a particular group of individuals who are similar

to that individual to whom the test is designed. Darwesh (2003:3) argues that in a NRT, scores are interpreted in relation to other scores so that it is possible to tell who is better or worse than whom. The designer of a NRT usually selects items that are supposed to be answered by about 50 percent of the population understudy and takes that as his norm. This is perhaps why the scores achieved can be typically represented in the shape of a bell called 'normal bell-shaped curve'. Look at Figure 1 taken from Bachman (1990:73)

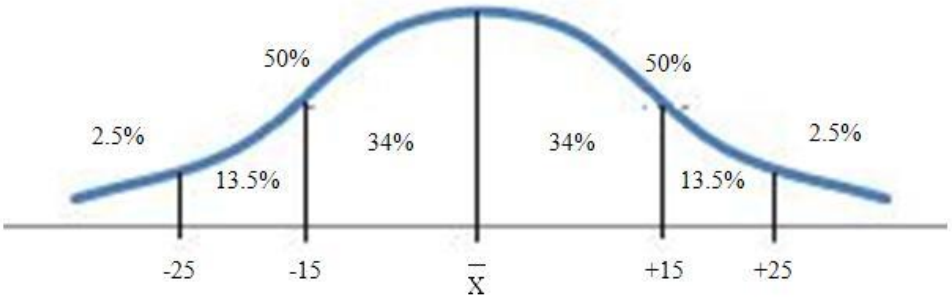

Figure 1: A bell-shaped curve of normal distribution of scores of a typical NRT

A NRT can tell a teacher or an educator how a particular student compares to a similar student on a given skill or knowledge, but it does not supply information about what that student can and cannot do (Popham, 1975). Scores on NRTs identify the student's ranking relative to a 'norm group'.

If the typical scores of a NRT do not provide information bout what a particular learner can and cannot do, then what do they include? Canter (1998) provides an answer for this important question by stating that they include percentiles, stanines and standard score.

2.1.1.1 Percentiles

A percentile is a score that indicates the rank of the student compared to others using a hypothetical group of 100 students. In using percentiles, one needs to be aware that the units are of unequal values. Therefore, one should not add percentile scores from one test to those derived from another to find a total or an average. This is so because each percentile is derived from raw scores using norms obtained from testing a population when the test was first developed.

2.1.1.2 Stanines

According to Canter (1998), stanines are essential groups of percentile ranks, with the entire group of scores divided into 9 parts, with the largest number of individuals falling in the middle stanines (3.7) and fewer students falling at the extremes.

2.1.1.3 Standard Scores

Standard scores are intended to show how a student compares with those of his own group. A standard score is derived from raw scores using the norming information gathered when the test was first developed. According to Jackson (1973:37), standard scores can be used to compare individuals from different grades or age groups for scores can be converted to the same numerical scale.

The researchers are not quite interested, here to give a full account of the statistical traps required for building or interpreting the scores of NRTs, but to show that these tests cannot be prepared by ordinary classroom teachers since they require complex statistical methods, and that they are used to compare learners using ‘normal’ or ‘bell-shaped’ figures. They are not indices of what an individual learner does or does not know which should be the ultimate aim of the teacher to check

the progress of his learners and to prepare the necessary remedial work.

2.1.2Criterion-Referenced Tests

CRTs fall in contrast with NRTs. In CRTs, the major aim is to produce a clear-cut description of what a learner's performance on a particular test actually means, i.e., to give description of what a learner can do or cannot do.

A CRT is the test which provides for translating the test scores into a kind of report about the behavior to be expected of a student with those scores or their relationship to a specific subject matter or skill. Therefore, the items a test constructor prepares should measure whether or not a learner has mastered a set of specific objectives. Emphasis, in this respect, should be on the use of behavioural objectives set for a particular skill or domain of knowledge (Darwesh, 2003:4). Perhaps, this is why Glaser (1963) insists that most tests and quizzes written by school teachers should be CRTs since these tests can provide very useful information about achievement and ranking. According to Wikipedia (2008), two-thirds of U.S. high school students are required to pass a CRT of high school graduation examinations. One high fixed score is set at a level adequate for university admission whether the high school graduate is college bound or not. Each state gives its own test and sets its own passing level which is a(cut score).In order to ensure quality in Iraq, our higher educational institutions should follow such systems and it is time that all Iraqi Educational Institutions should establish boards of examinations and assign to them the taking of all such important decisions of running the educational policy in the country. One might say that there are such educational institutions. But according to the researchers' point of view, even if there are such institutions, they are not given the

full right to issue important decisions and there are not enough members of qualified personnel to run these institutions.

However, one of the problems with CRTs is the meaning of 'criterion'. Some people believe that the term 'criterion' is not quite definite or clear since it refers to the score and the test at the same time. This common misconceptions springs from the fact that CRTs involve a 'cut score' where a test-taker passes if he exceeds the cut score and fails if he does not. The cut score is not the criterion. According to Bachman (1990:75), the criterion is the 'domain of content, or the 'level of ability'. "It is important to point out that it is this level of ability or domain content that constitutes the criterion". (ibid: 79)

As for Allam (1986:17), the source of misconception comes from the fact that some people believe that the 'criterion' is a 'standard' which equates CRTs with NRTs. Ventre (2000) backs this idea by arguing that because of this misconception, CRTs have also been called 'standard-based assessment' by some education agencies as students are assessed to 'standards' that define what they should know.

The researchers do not want to go deep in this controversy but would like to state that the real meaning of 'criterion' is to measure a learner's performance according to a comprehensive domain of knowledge and skills which are well specified in the form of behavioural objectives (Glaser, 1963). The degree of precision of specification of a special domain or skills depends largely on the content of that domain and on the range of complexity of the objectives required of these tests. This is, perhaps, why some test experts call these tests as 'objective-referenced tests' or 'domain-referenced tests'.

The major concern when constructing CRTs is that they are intended to represent the ability level or sample the content

domain to be rested. In fact, they should be sensitive to levels of ability or degrees of mastery of components of that content domain. This sensitivity is decided by the adopted' cut score'. Therefore, prior to the development of a CRT, there must be a specification of a level or domain content (Glaser, 1963).

2.1.3 Reliability and Validity of NRTs and CRTs

At the outset of tackling reliability and validity of these types of testing, the researchers would like to state that they do not have the intention of going deep into details of these two aspects of language testing because this topic has been fully exhausted by many people such as (Harris, 1967; Oller, 1979; Nitko, 1983; Oller, 1983; Bachman, 1990). The aim is to give a tentative idea concerning the tension that exists in the construction of theses two types of testing. Concerning reliability which is described as the stability of scores gained as a result of administrating a test, NRTs are usually described as being highly reliable but there are doubts concerning their validity as accurate measures of true language ability of individual learners since they stress taking the 'normative sample' as their criterion for judging the performance of an individual. On the other hand, the validity of CRTs, which is described as fitness for purpose (Darwesh, 1986), is considered as quite satisfactory. However, the calculation of their reliability is not quite straightforward. What is problematic, here, is the fact that there is no such thing as validity without reliability and that a test without validity does not worth anything, to anybody, at anytime, for any purpose. Therefore the information about the performance of any individual learner or a group of learners must be valid and reliable.

3.1 A Comparison between NRTs and CRTs

In order to draw a comparison between NRTs and CRTs to filful the second aim of this study, Bachman (1990:75) argues that there are two important distinctions that should be stressed, i.e., design, construction and development and the interpretation of the scales they yield. As for Popham (1975), he presents his comparison in four dimensions: purpose, content, item characteristics and score interpretation. This study is going to follow Popham (ibid) in presenting the comparison. It might be useful if the tension of validity-reliability between the two types of testing is added to this comparison.

3.1.1 Concerning purpose, As for NRTs, they are intended to rank each learner in relation to the performance of other learners who take these tests in broad sense of knowledge. The idea behind this is to discriminate between those who have done well or poorly on the adopted test(s). They usually intended to assist institutions in the selection, placement, and readiness for instruction for a particular programme of learning. Therefore they are not intrinsically intended for checking learners' progress in a particular programme of instructions

CRTs, on the other hand, are intended to determine whether each learner has achieved specific ability or skill. They also aim at finding out how much learners know at the beginning of the instruction and after it has finished.

CRTs are also intended to evaluate the learner's mastery of the objectives of a given course of instruction and not interested in the comparison of individual learners or groups. As a result, teachers usually use them as end-of-unit tests of the prescribed textbooks and to upgrade learners to a higher level of instruction in addition to gauging the progress of learning of the individual learner.

3.1.2 As for content, NRTs are usually intended to measure specific skills which make up a designated syllabus. Such skills are identified by educators, teachers and syllabus designers. Each skill is translated into instructional objectives. The items of these tests are to be selected according to how well they discriminate individuals from one another, or a whole group from another group. They are also selected according to how adequately they represent the ability level which the test(s) is supposed to measure. The linguistic ability of the learner is usually represented in the form of the four skills of language, i.e., listening, speaking, reading and writing. Each individual skill is subdivided into a number of sub-skills which are tested at a rate of 'one language point at a time'. This discrete-point approach to language testing has been criticized by Oller (1979) and Bachman (1990).

CRTs are designed to measure the achievement of an individual learner with reference to well-defined domains or skills which are predefined by the content of the tested domain or skill and not by educators, teachers, or the like people. Discriminations among individual is not sought for in their design though it is possible to carry them out after calculating the results of the test(s), (Bachman and Clark, 1987). Bachman (1990:74) argues that CRTs,"are based on a fixed standard content' which does not vary from one test format to another. Such content can be based on a theory of language proficiency rather than on individual subs-kills. The mastery of each skill or domain is expressed as the goals of any teaching and testing process.

3.1.3 Concerning the general characteristics of test items, NRTs dictate that individual items of a test should vary in their level of difficulty so as to leave room for the examiner(s) to discriminate high or low achievement of the testees, (Deale

1975:158). They usually test an individual language item not more than twice because they adopt a discreet-point approach to language testing. This approach necessitates the writing of many test items to test an individual language skill or a sub-skill. Therefore if test designers want to lessen the guessing factor of the test and test each language item more than twice, then they will end up with a very long test that cannot be practicable to administer, (Darwesh and Al-Jarah, 1989:21).

CRT items are parallel in their difficulty level since their primary aim is to reveal the language ability of each individual learner apart from the performance of other individuals. CRTs usually aim at testing each language skill or a domain of content, at least, four times so as to arrive at a very good sampling adequacy of testees' performance and to lessen the guessing factor as far as possible ,(Bachman, 1990).

3.1.4 Score interpretation is another criterion to be discussed in the comparison of the two approaches to language testing. At pages (4-7), some of the statistical methods that are used in the interpretation of test results of NRTs have been tackled; therefore there is no need to re-mention them here again and concentrate on the fact that the interpretation of the results of many NRTs are of very limited value when used with a learner or a group of learners who are intrinsically different from the 'norm group' from which the interpretation of the results of the test scores are originally derived. In such cases, the results obtained from such learners may not reflect their original language ability since all the assumptions which are concerning their performance are not applicable. This is one of the most serious criticisms directed against NRTs.

In a CRT, the score of each individual is compared with preset assumptions for an acceptable standard of achievement usually

predetermined by a 'cut score'. Unlike that of the NRTs, in CRTs the performance of other individuals is irrelevant because the principal aim is to know the performance of each individual apart from that of the group. Learners' scores are usually given in the form of percentages to reflect the level of mastery of each learner for the objectives of the course of instruction. This should not be confused with 'percentiles' which are used in ranking of learners in NRTs. The 'percent correct' refers to what a particular learner has gained as correct out of the items that are set to cover a specific domain or skill with a 'cut score' which determines the 'pass/fail' level.

3.1.5 The validity-reliability tension:

The validity-reliability tension of both NRTs and CRTs has already been hinted at (see 2.1.3). What is of interest here is that CRTs are accused of stressing the importance of validity at the expense of reliability. In defense of this accusation, Bachman (1990:211) argues in CRTs that:

> It is not the case that reliability is of no concern in {CRTs}.On the contrary, consistency, stability, and equivalence are equally important for CRTs. However, they take on different aspects…and therefore require different approaches to both estimation and interpretation.

The validity of most of the NRTs is highly questionable since they cannot measure progress of the population of a whole, only where individuals fall within the whole, (Wikipedia,2008).

4.1 Conclusions and Recommendations

4.1.1 Conclusions

As a result of surveying and comparing the two types of approaches to language testing, the researchers have come to the following conclusions:

1. NRTs are only designed and developed to measure the progress of individuals, when those individuals fall within a whole which constitutes the 'normative sample'. They are intended to maximize distinctions among individual test-takers, i.e., items will be selected according to how well they discriminate individuals who do well on the test as a whole from those who do poorly.

2. Standardized tests cannot be adopted right away by the classroom teacher for the assessment of his students' performance since they assume a particular linguistic level predetermined by the 'normative sample'.This level might not coincide with the level of his students. Furthermore, standardized tests require pre-testing and post-testing processes that fall, most of the time, beyond the professional ability of the classroom teacher.

3. CRTs are designed to be representative of levels of linguistic ability or domains of content. The items are selected according to how adequately they represent the levels of ability or domains of content. In this, CR approach to language testing can be applied to the development of language proficiency of learners as well as to the evaluation of instructional programmes.

4. CRTs ignore the principle of item difficulty which may sometimes cause problems and learners are usually met with items that fall beyond their language ability. By contrast, NRTs do not face such a problem because they do not seek to enforce any expectations of what a learner should know.

5. Guessing factor and the probability of getting the correct answer of an item by mere chance are higher with NRTs than with CRTs since the former depend heavily on objective tests using discrete-point approach to language testing, i.e., multiple-choice items, true/false items, etc.

6. Sampling adequacy is more secured with CRTS than with NRTs since test items may have the chance of being tested for four times. This ensures the 'internal consistency' of the whole test and even lessens answering by mere chance.
7. The reliability of NRTs is highly dependable and systematically calculated. Its estimate procedures are well defined and statistically possible. However, CRTs do not ignore reliability and they do have their ways of calculating reliability.

4.1.2 Recommendations

Depending on the above conclusions, the researchers would like to recommend:

1. The use of CRTs for the assessment of the achievement of learners inside language classrooms since they provide useful information about actual achievement, progress in the course of instruction and relative ranking of a learner.
2. The reliance on only tests and scores is not always adequate to give a clear-cut evaluation of learners' performance. Therefore the researchers recommend the use of other types of information that can describe the true language ability of learners such as report writing, transforming scores into grades, etc.
3. Teachers of English should receive post-serves training in language testing to update their own knowledge particularly in new trends or when a new syllabus is introduced as is the case in Iraq nowadays.
4. The abundance of using standardized tests or even to mimic their techniques is simply dangerous because they require pre and post testing before one can adopt them. They are not even suitable to check progress of learners in a particular programme of instruction.
5. The adoption of CRTs for the evaluation of students' performance inside classrooms since they are valid, do not need pre or post-testing and enjoy reliability as well.

Bibliography

- Allam, s. (1986) Modem Development in Psychological and Educational Measurement, University of Kuwait, Al-Qabas Printing House.
- Bachman, L. F. (1990) Fundamental Consideration in Language Testing, Toronto, OUP.
- Canter, A. (1998) Understanding Test Scores, Network for Instructional TV Inc.
- Cronback, L. J., (1970) Essential Psychology Testing, New York, Harper and Row.
- Darwesh, A-J., (1989) The Techniques by which Teacher-Made Tests are Evaluated, Al-Mu'ullim Al-Jadid 2, vol 43 : 9-10.
- ---------- (2003) Testing in 'Rafidain English Course for Iraq' a Publication of the Ministry of Education.
- Deale, R.(1975) Assessment and Testing in the Secondary School, London, Evans Methuen Educational.
- Glaser, A., (1963) Instructional Technology and Measurement of Learning Outcome, American Psychology 18 510-e22.
- Jackson, S. (1973) A Teacher's Guide to Tests and Testing, London, Longman.
- Nitko, A., (1996) Educational Assessment of Students, Englewood, .
- Popham, J. W., (1975) Educational Evaluation, New Jersey, Prentice-Hall Int.
- Sanders, W. and S. Horn (1995) Educational Assessment Research, Englewood Cliff, Prentice-Hall Inc.
- Venture, M. (2000) Assessing the Assessment of Outcomes Based Education, Cape Town, and Conference Papers 2000 art. 3-9.
- Weir, C.(1988).Communicative Language Testing, Oxford, OUP.
- Wikipedia (2008) free encyclopedia, Criterion-referenced Tests, retrieved from (htto://www.citrus.kcusd.com/instruction.htm).