CHAPTER - 2
CRITERION-REFERENCED TEST THEORY

2.1 Introduction

2.2 Nature of CRT

2.3 NRT and CRT

2.4 The Developmental Aspects of CRT
   2.4.1 Preliminary Consideration
   2.4.2 Domain Specification
   2.4.3 Item Writing
   2.4.4 Logical Review of Test Items
   2.4.5 Revision of Test Items
   2.4.6 Empirical Review of Test Items
   2.4.7 Revision of Test Items
   2.4.8 Test Assembly
   2.4.9 Validity
   2.4.10 Reliability
   2.4.11 Selection of Standard
   2.4.12 Test Manual

FootNotes
CHAPTER - 2

CRITERION-REFERENCED TEST THEORY

2.1 Introduction

In the field of education, learning is a very important process. This process is having some objectives. The teacher and the pupil are very important factors of the teaching learning process.

At the end of teaching learning process, it is necessary to evaluate the knowledge, skill, aptitude and educational objectives whether they are fulfilled or not. For this purpose different kinds of tests have been adopted by researchers and teachers. Various experiments such as Programmed Instruction, Mastery Learning, Personalized Instruction, Project Plan, Micro Teaching etc. are being tried for the improvement of education.

Today, it is the main task of the institutions related with education to improve or uplift the level of education. As a result of this, various experiments are carried out by the researchers to measure the achievement of the examinees. For this, different kinds of tests have been constructed by the researchers.

Traditional achievement tests viz. NRTs are seen as less than ideal for providing the desired kind of test score information. It is felt that these tests are much less appropriate when the question asked related to what a student can and cannot do with respect to a well-defined domain of behavior measuring are objective or skill.

M.B.Buch (editor) writes in this connection:

"To add another characteristic to the Indian researches in the area of Educational Evaluational and Examinations, it may be stated that most of our studies are conducted following the traditional steps of tool construction within the frame work of norm referenced testing. New
concepts like 'criterion-referenced testing' and 'mastery learning' have to be tried in the tool construction.”¹

As an alternative criterion-referenced tests are developed to meet the testing and measurement requirements in objective-based instructional programmes. Instead of Norm-referenced Tests (NRTs), Criterion-referenced Tests (CRTs) have been adopted now.

R.A.Berk writes:

"Among the many developments in educational measurement over the part 50 years, the emergence of criterion-referenced testing has been one of the most dramatic. An emphasis on mastery-proficiency-competency has permeated all levels of education, occupational licensure and certification (particularly in medicine and the allied health fields), and business and industry, commercial test publishers have committed substantial resources to the production of criterion-referenced testing systems and diagnostic-prescriptive instructional packages for decision making at the classroom level.”²

The research into and application of this concept, however did not begin formally until the 1960s.

A Brief History: (By R.A.Berk)

1960s:

The term criterion-referenced was first coined in 1962 by Glaser and Klaus (1962)³. One of the first articles on the topic of criterion-referenced testing appeared in the 'American Psychologist' (Glaser 1963)⁴. Over 700 papers on the topic have been published since then and the scope and direction of educational testing has been changed dramatically. Glaser was interested in assessment methods that would provide necessary information for making a
number of individual and programmatic decisions arising in connection with specific objectives or competencies. Norm-referenced tests were seen as limited in terms of providing the desired kinds of information. For the remainder of the 1960s there was a sparcity of major research activity on the topic until publication of the article by Popham and Husek in 1969, which seemed to provide the stimulation needed to arouse the measurement community. Their article amplified Glaser's distinction between criterion-referenced and norm-referenced measurement and also discussed the advantages and disadvantages of the two approaches in instructional decision making.

1970s. Just prior to 1978, Hambleton, Swaminathan, Algina and Coulson (1978) estimated that more than 600 papers had been written on criterion referenced testing. Many of these papers appeared in a variety of journals, monographs, and books, but a proportionately greater number were unpublished in the form of conference presentations and reports by school districts. The researchers gave different titles to these reports. Although some researchers consistently used the label “Criterion-referenced test” it is not uncommon to find the terms domain-referenced test, objective-referenced test, competency-based test, proficiency test, and mastery test in the literature.

The research literature in the 1970s indicated a need for synthesis and direction. During 1974 a monograph was also published on the technical problems in criterion-referenced measurement which was edited by Chester W. Harris, Marvin C. Alkin, and W. James Popham (1974). In 1977 Victor R. Martuza produced the first measurement textbook at integrated norm-referenced and criterion-referenced test construction procedures.

There was an edited monograph on criterion-based measurement by Bunda and Sanders (1979), a monograph on criterion-referenced testing by Brown (1980) edited volumes by Jaeger and Tittle (1980) and Berk (1980) based on the aforementioned conferences, and a special 1980 issue of 'Applied
Psychological Measurement: examining the criterion-referenced testing technology which was guest edited by Ronald K. Hambleton.

According to Hoko J. Aaron:

"Criterion-Referenced Tests (CRTs) provide teachers with skill specific information concerning their students' academic development and thus enable teachers to effectively map out detailed instructional plans."  


"The CRT is an instructional strategy. This strategy assists an individual to learn something better than he would have learned it by himself."  

2.2 Nature of CRT

At least 57 definitions are available for criterion-referenced test in the CRT literature (Grey 1978 and Nitko 1980).

These definitions are given by different scholars from time to time to explain the nature and uses of CRTs. Some of these explanations are as below:

1. According to W.J. Popham (1978):

"A criterion-referenced test is used to ascertain an individual status with respect to a well-defined behavior domain."  


"Criterion-referenced tests are constructed to permit the interpretation of examinee test performance in relation to a set of well-defined competencies."  


"A CRT is one that is compared of items keyed to a set of behavioural objectives."

"A CRT is one that is deliberately constructed to yield measurement that are directly interpretable in terms of specified performance standards."\(^{18}\)

Hambleton (1982) modified Popham's definition of CRT and that modified definition is as below:

"A criterion-referenced test is constructed to assess the performance level of examinee in relation to a set of well-defined objectives (or competencies)."\(^{19}\)

The definition that is given by Popham is the one which is most widely used. Criterion-referenced testing, is alternately referred to as domain-referenced testing, objective-referenced testing, basic skills testing, competency testing, or performance based testing.

The main concern in development of CRT is to obtain rigorous and precise domain specification to maximize interpretability of domain score and estimation of an individual domain score. The allocation of examinees to mastery-non mastery states is a noteworthy advancement of CRT. The test in which the cut off score is determined is called mastery test. Moreover, CRT depends on absolute standard of quality which involves the conception of change.

Hambleton gives more points about CRTs as below:

"Three additional points about criterion-referenced tests deserve mention: (a) the number of competencies measured by a criterion-referenced test will (in general) vary from one test to the next. (b) the number of test items measuring each competency and the value of the minimum standard will (in general) vary from one competency to the next, and (c) a common method for making mastery-nonmastery decisions involves the comparison of examinee percent (or
proportion-correct) scores on competencies to the corresponding minimum standards. With respect to (c) when an examinee’s percent score is equal to or greater than the standard, the examinee is assumed to be a master” (M) otherwise the examinee is assumed to be a “nonmaster” (NM). There are however more complex decision-marking models (for review see van der Linden 1980).

It is common to see terms like criterion-referenced tests, domain-referenced tests (see Domain-referenced Tests) and objectives-referenced tests in the psychometric literature. Popham’s definition for a criterion-referenced test is similar to one Millman and others proposed for a domain-referenced test. There are no essential differences between the two if Popham’s definition for a criterion-referenced test” is a descriptive one and therefore it is less likely to be misunderstood than the term “criterion-referenced test.” One reason for continuing to use the term “criterion-referenced test.” even though it is less descriptive and its definition has become muddled in the psychometric literature, is that there is considerable public support in the United States for “criterion-referenced tests.” It would seem to be a waste of valuable time to mount a campaign for a new term.

Objectives-referenced tests consist of items that are matched to objectives. The principal difference between criterion-referenced tests and objectives-referenced tests is that in a criterion-referenced test, items are organized into clusters with each cluster serving (usually) as a representative set of items from a clearly defined content domain measuring an objective, while with an objectives-referenced test, no clear domain of content is specified for an objective, and items are not considered to be representative of any content domain. Therefore, interpretations of examinee performance on objectives-referenced tests should be limited to the particular items on the test.”

30
2.3 NRT and CRT

The traditional NRT held away till CRT come into prominence in recent times. Yet both these testing measurements have their own importance and utility. It becomes necessary to note the important points of differences between these two type of measurement instruments.

1 According to Popham (1981) the difference between NRT and CRT is as below:

"The fundamental distinction between NRT and CRT is basically based on the manner in which one interprets the results of an examinee’s test performance. In case of NRT, one interpretes some one’s test performance according to the performance of others, in case of CRT one interpretes some one’s test performance in relation to a well-defined class of knowledge, skills, attitude and the like. In very real sense, interpretations are made relatively for NRTs and absolutely for CRTs." 21

2 Popham further comments in his article (1993) as below:

"During the late 1960s and early 1970s, criterion-referenced measurement was viewed by most psychometric specialists as an unwelcome intruder on the traditional educational measurement scene. Norm-referenced tests, after all, had been around for several decades and were highly regarded. From a measurement traditionalist’s perspective, these upstart criterion-referenced outriders were treading on sacred testing terrain. Moreover, many criterion-referenced advocates cast negative aspersion on norm-referenced measurement’s relative assessment strategies and proclaimed the virtues of criterion-referenced measurement’s absolute assessment strategies."
It's not that norm-referenced tests were totally rejected. Most advocates of criterion-referenced measurement endorsed the utility of norm-referenced tests in fixed quota settings. However, for purposes of program evaluation or instruction, norm-referenced tests were pretty solidly scorned by criterion-referenced enthusiasts. 22

Hambleton points out the differences in test purpose, area of content specificity, are of test development and test score generalizability between NRT and CRT as follows:

3 A norm-referenced test is constructed specifically to facilitate comparison among examinees in the content area measured by the test ... Criterion-referenced tests on the other hand, are constructed to assess examinee performance in relation to a set of competencies....

4 It is common for designers of both test types to prepare test blueprints or tables of specifications. It is even possible that norm-referenced test designers will prepare behavioural objectives. But, criterion-referenced test designers must (typically) prepare considerably more detailed content specifications than provided by behavioural objectives to ensure that criterion-referenced test scores can be interpreted in the intended way....

5 With NRTs, item statistics (difficulty and discrimination) serve an important role in item selection. In contrast, criterion-referenced test items are only deleted from the pools of test items measuring competencies when it is determined that they violate the content specifications....

6 The fourth and final major area of difference between CRTs and NRTs is test score generalizability. Seldom is there interest in making generalization from norm-referenced achievement test scores. In contrast, score generalizability is usually of interest with CRTs....
Livingston (1972) cited by Richard et.al. (1972) observed as follows in this connection:

“When we use norm-referenced measures, we want to know how far a student’s score deviates from the group mean. When we use criterion-referenced measures, we want to know how far his score deviates from a fixed standard, the criterion.”

Carver (1974) draws a line of demarkation between NRT and CRT when he says,

“The tests that have focused on measuring between individual differences have been called psychometric tests. Therefore, this dimension of tests will be called the psychometric dimension... The other dimension of tests will be called the edumetric dimension. A test may be evaluated in terms of its edumetric properties, that is, the extent to which it reflects the within individual growth that traditionally has been of primary interest to educational testing.”

A few other points of differences between the two tests are also noteworthy. They are as follows:

NRTs are designed to produce test score suitable for ranking individuals on the ability measured by the test. CRTs are specially designed to meet the measurement needs of the new instructional models.

NRTs are particularly useful in situations where there is interest in “fixed quota” selection on some ability continuum. CRTs involve what Cronbach and Glaser (1965) would call a “quota free” selection.

In NRTs emphasis is on measurement of achievement only but in CRTs the emphasis is on improvement of student’s achievement. Where as no criterion is necessary in NRTs, the specification of criterion level of
minimum acceptable performance for each objective is a prerequisite in CRTs.

R.K. Hambleton also gives following points for the difference of NRT and CRT as below:

"Norms-referenced achievement tests are needed to provide reliable and valid normative scores for comparing examinees. Criterion-referenced achievement tests are needed to facilitate the interpretation of examinee performance in relation to well-defined competencies."

Although the differences between norm-referenced tests and criterion-referenced tests are substantial, the two kinds of tests share many features. In fact, it would be a rare individual who could distinguish between them from looking at the test booklets alone.

There are a number of important differences, however, between them. The first difference is test purpose. A norm-referenced test is constructed specifically to facilitate comparisons among examinees in the content area measured by the test. It is common to use age-, percentile, and standard-score norms to accomplish the test's purpose (see Test Norms). Since test items are (or can be) referenced to competencies, criterion-referenced score interpretations (or, more correctly, objectives-referenced score interpretations) are possible but are typically limited in value because of the (usually) small number of test items measuring any competency in the test.

Criterion-referenced tests on the other hand are constructed to assess examinee performance in relation to a set of competencies. Scores may be used (a) to describe examinee performance, (b) to make mastery-nonmastery decisions, and (c) to evaluate program effectiveness. Scores can be used to compare examinees but comparisons may have relatively low reliability if score distributions are homogeneous.
The second difference is in the area of content specificity. It is common for designers of both test types to prepare test blueprints or tables of specifications. It is even possible that norms-referenced test designers will prepare behavioral objectives. But, criterion-referenced test designers must (typically) prepare considerably more detailed content specifications than provided by behavioral objectives to ensure that criterion-referenced test scores can be interpreted in the intended way. This point will be considered further in the next section. Thus, with respect to content specifications, the difference between the two types is in the degree to which test content must be specified.

The third difference is in the area of test development. With norm-referenced tests, item statistics (difficulty and discrimination indices) serve an important role in item selection. In general, items of moderate difficulty (p-values in the range 0.30 to 0.70) and high discriminating power (point biserial correlations over 0.30) are likely to be selected for a test because they contribute substantially to test score variance. Test reliability and validity will, generally, be higher when test score variance is increased. In contrast, criterion-referenced test items are only deleted from the pools of test items measuring competencies when it is determined that they violate the content specifications or standard principles of item writing, or if the available item statistics reveal serious noncorrectable flaws. Item statistics can be used to construct parallel forms of a criterion-referenced test or to produce a test to optimally discriminate between masters and nonmasters in the region of a minimum standard of performance on the test score scale.

The fourth and final major area of difference between criterion-referenced tests and norms-referenced tests is test score generalizability. Seldom is there interest in making generalizations from norm-referenced achievement test scores. The basis for score interpretations is the performance of some reference group. In contrast, score generalizability is usually of interest with criterion-referenced tests. Seldom is there interest in the performance of examinees on specific sets
of test items. When clearly specified competency statements are available and assuming test items are representative of the content domains from which they are drawn, examinee test performance can be generalized to performance in the larger domains of content defining the competencies. It is this type of interpretation which is (usually) of interest to criterion-referenced test users.\textsuperscript{25}

These points of differences between NRT and CRT point out the reason why there is greater inclination towards CRT than towards NRT.

2.4 The Developmental Aspects of CRT

As CRT is a recent innovation particularly in India, a brief discussion, in due order, of theoretical aspects regarding the development and validation steps of CRT becomes necessary for the better understanding of the subject. It is from this point of view that each of the twelve steps advocated by Hambleton\textsuperscript{26} (1978, 1982, 1985) and explained in detail by Modi\textsuperscript{27} (1986, p.1-20) is briefly described one by one:

2.4.1 Preliminary Considerations

It is a common knowledge that a test developer has to give thought to certain points concerning a test before he/she starts to develop a test and to have a clear picture of the following aspects:

(a) test purpose;

(b) specification of content area;

(c) identification of the groups to be measured and

(d) identification of qualified staff.

The purpose of a test influences the appropriateness of the breadth of a domains. If the purpose is to provide feedback, the coverage of all the units becomes necessary; if the purpose is to find out causes of recurring learning difficulties (diagnosis), the selection of content is based on common sources of
learning errors and if the purpose is to assign grade or certify mastery, the coverage of the content is wider and items have a wide range of difficulty.

From the above-mentioned four aspects, (b) and (c) are self-explanatory.

The services of the subject-matter experts and evaluators become inevitable at the stage of domain specification, logical review and content validity. A test developer has to plan well in advance to secure the services of such experts who are well-versed in the subject and have a considerable experience of this nature.

2.4.2 Domain Specification

Domain specifications are important new development in CR Testing (Baker 1974; Millman 1974; Popham 1978, cited by Hambleton 1982). Hambleton's (1980) following remark in this connection is apt and appropriate.

"Well-defined objectives make the task of writing test items easier and improve the quality of test score interpretations. Writing is easier because appropriate content is spelled out. The quality of test score interpretations is improved because of the clarity of the content or behavior domains to which test scores are referenced."

A domain is explained by Baker (1985) in the following words,

"Domain refers to the specifically circumscribed universe from which performance is sample and to which performance is expected to generalize.

The above mentioned valuable observations regarding domain specification will prove to be torch-bearers for a research worker in this particular field."

The real task of test development begins after giving proper thought to the remarks made by Hambleton and Baker.
At this stage, a researcher has to obtain a completely clear picture of the content. It is at this stage that a researcher makes an analysis covering all the aspects of content in the form of skills and/or competencies. Teaching of which topic under which conditions brings about the expected changes in behavior of an examinee is described by each domain. After the specification of domain is done and written in objective form, it becomes necessary to get the objectives examined by the subject matter experts from the print of view of their appropriateness for the level for which they are meant.

Using as a basis the work of Popham (1978), Hambleton (1982) suggested that a domain specification might be divided into four parts:

(a) Description—a short, concise statement of the content and/or behaviors covered by the competency.

(b) Sample directions and test item—an example of the test directions and a model test item to measure the competency.

(c) Content limits—a detailed description of both the content and/or behaviors measured by the competency, as well as the structure and content of the item pool. (This section should be so clear that items may be divided by reviewers into those items that meet the specifications and those items that do not.) Sometimes clarity is enhanced by also specifying areas which are not included in the content domain description.

(d) Response limits—a description of the kind of incorrect answer choices which must be prepared. The structure and content of the incorrect answers should be stated in as much detail as possible.

2.4.3 Items writing

Domain specifications lead a researcher to generate items. There is a subjective element in item writing in NRT. Some of the leading pioneers in the
field of CRT tried to make the domain specified and some others concentrated their attention on finding out item writing techniques which were capable of reducing subjectivity in item writing. Item writing in NRT was considered more as an art than as science before the theory proposed by Roid and Haladyna (1982) came into light. As this consideration of item writing is more of art and less of science, it becomes necessary to remove or reduce the subjectivity in item writing because the quality of a test is dependent on the quality of items.

Roid and Haladyna (1982) suggest a solution for the removal of deficiencies in conventional methods in the following way, by providing,

1. Item-writing methods based on logical and precisely defined relationship between the instructions of the unit given and the test items written.

2. Item-writing methods defined by a set of operations open to public inspection rather than defined by the private, intuitive standards of the item writer.

3. Item-writing methods that are specific and public so that they can be replicated by many test developers, and not just the original item writer.

Item writing methods of the kind described will produce tests that are more scientific, thereby contributing to the advancement of instructional research, educational evaluation and the use of test data in the formation of public policy.  

Herman (1985) emphasizes the same point when he says,

"Item writing techniques provide rules and prescriptions for constructing sound test items, items that measure what they are intended to measure."
Roid and Haladyna (1982) have discussed six different item-writing techniques which are as follows:

(a) Item-writing for prose learning
(b) The mapping sentence method/Facet Design
(c) Domain based concept testing
(d) Logical-operations for generating intended questions
(e) Item-forms
(f) Instructional Quality Inventory (IQI)


2.4.4 Logical Review of Items

Every item written for a test is subject to certain forms of analysis to determine its adequacy to meet the intended purpose. Traditional item analysis was limited to estimating each item difficulty value and discrimination indices. With the advent of CR Testing, the need for two new types of item reviews has emerged: logical and empirical.

The logical review refers to the logical relatedness of items to their instructional intent. Logical review establishes content validity of CRT items. Hambleton (1981) mentioned a judgmental approach which is used to conduct logical review. That approach required content specialists to provide judgment of test items. Several procedures suggested by Hambleton et al. (1982) are as follows:

(a) Item-objective congruence index (IOCI)
(b) Matching task
(c) Rating of appropriateness
(d) Cronbach’s duplication experiment
Over and above these procedures, several other educationists like Berk (1978, 1980); Brown (1980, chap-4); Haladyna and Roid (1981, 1982) as well as Roid and Haladyna (1982, chap-13) cited by Berk 34 (1984, p-97) have pointed out form time to time different procedures for logical item review.

The representativeness of the items measuring each objective is determined by adopting any one of the above mentioned procedures.

2.4.5 Revision of Test Items

A test developer has to revise the test items if they are found to be deficient in any way by the subject-matter experts. In a CRT, the deficient items are not to be discarded but are to be revised as the items are generated on the basis of well-defined behavioural domains.

2.4.6 Empirical Review of Items

"In spite of the items being generated on the basis of rigid domain specification by employing the most sophisticated item-writing technology and inspite of their being logically reviewed still however the experience of experts in the field indicates that there will always be some objective element in formation of objectives and therefore in generation of items. Hence, there is a room for some kind of item analysis" 35 (Wood 1985).

A test developer has to carry out empirical review through the measurement of the examinees responses. There are three main approaches for the empirical review:

(a) to examine the instructional sensitivity of the items;

(b) to examine the difficulty value and discrimination value of items according to the traditional method and

(c) to examine the quality of items through item characteristic curve theory.
A test developer has to adopt any one of these three approaches. The empirical review provides useful information for detecting 'bad' items. The aim of empirical review in CRT is to find out items which are not sensitive to instruction. The data obtained from the examinees response are used to evaluate the effectiveness of items in terms of whether they function in the manner in which they were intended.

2.4.7 Revision of Test items

After reviewing the items empirically, it becomes necessary to revise the item/items if found deficient in their difficulty value and in language. Such a revision enhances the quality of items.

2.4.8 Test Assembly

This step consists of four aspects of the final form of the test. The aspects are:

(i) Parallel forms;
(ii) Test Length;
(iii) Item selection and
(iv) Miscellaneous

After the above mentioned two reviews, the final form of the test is prepared at this step of test assembly.

(i) In CRT, pre-post tests are ordinarilly administered to the groups for establishing the validation procedure.

(ii) Several researches are conducted on the subject of test length. It can be said on the strength of the experiments conducted by Hambleton et.al.(1976) that there should be at least eight items per domain and greater length of the test improves the goodness of fit to as modest extent.
(iii) The item selection is the heart of a CRT. In a CRT, it is to be decided what an examinee knows and does not know through the responses from him/her. A CRT domain always shows the homogeneous skills and competencies according to the nature of content. The items are generated on that domain. Consequently, if the items are representative of their respective domained, it can be stated with certainly what an examines knows and does not know. Two methods are prevalent for item selection for a CRT.

(a) Random sampling model and

(b) Classical model

A test developer has to adopt one out of these two methods/models.

(iv) After the preparation of a final form including general instructions in which examinees should answer the items as well as their personal data, the pilot test is administered. The pilot test examines several factors simultaneously. These factors are as follows:

(a) the general direction for the whole test;

(b) the general direction for the subtest;

(c) the duration of the period for the test;

(d) the difficulty facing the examinees in giving the responses to each item and

(e) whether the information expected to be obtained by the administration of the pilot test is available or not.

4 Again, the following three factors should also be considered:

(a) the schools where the final form of the test is to be administered;

(b) the students to whom the final form of the test is to be administered and
the physical facilities available at the schools where the final form of the test is to be administered.

Then in any of the above mentioned aspects, the revision, if found necessary, should be made on the strength of information available from the administration of pilot test.

2.4.9 Validity

The worthiness or unworthiness of any test depends upon its validity and reliability. Validity and reliability are established on the basis of the scores obtained by the examinees at the final form. Validity refers to appropriateness of particular uses of scores obtained from a test and not to test itself (Linn 1980). The evidence of validity is much needed for the intended interpretation and use of any set of test scores. It is mentioned by American Psychological Association (APA) that the questions of validity are questions of what may properly be inferred from a test score. Hambleton (1984) also observes that the validity of any uses of the test scores will need to be established within a general framework that three components: content, construct and criterion related validity. According to him, this is consistent with the new technical standards. With reference to construct validity.

Messick (1975) remarks that,

"Content validity is not sufficient because one is seldom interested in the performance defined by the domain of tasks but, rather, in the capabilities underlying this performance. Investigating this relation is construct validation." 32

Messick (1975) wrote:

"The major problem... is that content validity... is focused upon test forms rather than test scores, upon instruments rather than measurements. Inferences in educational and psychological measurement are made from
scores, and scores are a function of subject responses. Any concept of validity of measurement must include reference to empirical consistency. Content coverage is an important consideration in test construction and interpretation, to be sure, but in itself it does not provide validity. Messick defines construct validation as under:

"It is the process of marshalling evidence in the form of empirical relations to support the evidence that an observed response consistency has particular meaning."

Linn (1979) made a similar point

Questions of validity are questions for the soundness of the interpretations of a measure... Thus, it is the interpretation rather than the measure that is validated. Measurement results may have many interpretations which differ in their degree of validity and in the type of evidence required for the validation process.

The content validity is determined by judging the representativeness of the test items of a specified domain of content. Two requirements for assessing content validity are (1) a clear statement of the content domain and (2) details about the sampling plan used in item selection. Note that when test scores are reported at the objective level the content validity of each sample of items measuring an objective must be assessed.

It is usually desirable to have items in a criterion-referenced test that are representative of the content specified in a domain specification. In fact, the set of items should be broadly representative of the content domain of interest. However, to date only in some highly special cases has it been possible to specify completely a pool of valid test items. For example, there have been some successes in the areas of reading, mathematics, and spelling. But the examples are far removed from the more complex content domains in the areas of geography, world history and English.
Descriptions of content domains can range from broad areas of relevant content to tightly defined domains where every appropriate test item in the domain could be delineated. The former is unacceptable for criterion-referenced tests and the latter is impractical in most instances. Somewhere in between falls Popham's notion of a domain specification, which is both more reasonable and more practical than other methods presently available for specifying domains of content. If relevant content is described clearly, then it will be obvious to content specialists what domain a set of items is intended to represent. Categories or multidimensional classification schemes can often be developed to delineate further the content in a domain specification, and then content specialists can make judgements about the representativeness of the selected test items. When representativeness has not been achieved, new test items can often be added to a test and perhaps others can be removed.

Another procedure for assessing content validity is to carry out Cronbach's duplication experiment. This experiment requires two teams of equally competent item writers and reviewers to work independently in developing a criterion-referenced test.

Cronbach (1971) stated that

"They would be aided by the same definition of relevant content, sampling rules, instructions to reviewers, and specifications for tryout and interpretation of the data." 39

4 Construct Validity

Messick (1975) offered a useful definition of construct validation:

"Construct validation is the process of marshaling evidence in the form of theoretically evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning." 40
Construction validation studies have not been common in criterion-referenced measurement. This may be because criterion-referenced test score distribution are often homogeneous (for example, it often happens that before instruction most individuals do poorly on a test and after instruction most individuals do well). Correlational methods do not work very well with homogeneous score distributions because of problems due to score range restrictions.

But as Messick (1975) has noted.

"Construction validation is by no means limited to correlation coefficients, even though it may seem that way from the prevalence of correlation matrices, internal consistency indices, and factor analysis".

According to Ronald K. Hambleton:

"Construct validation studies should begin with a statement of the purposed use of the test scores, A clearly stated use will provide direction for the kind of evidence that is worth collecting".

Cronbach (1971) Stated –

"Investigations to be used for construct validation, then, should be purposeful rather than "haphazard"

CRT literature is rich with a number of methods for establishing the validity. Hambleton, in his article contributed to "A Guide to Criterion-Referenced Test Construction” edited by Berk (1984), discussed the following five approaches for establishing validity:

(a) Intra-objective approach;
(b) Inter-objective approach;
(c) Criterion-related approach;
(d) Experimental approach and
(e) Multitrait-multimethod approach
A test developer has to adopt a method or methods from one or more of these five approaches suitable to the intended purpose.

2.4.10 Reliability

The traditional co-relational estimates of reliability are quite inappropriate in case of CRT. Reliability indices for CRTs are identified and grouped into following three categories by Berk (1984):

(a) threshold loss function;
(b) square error loss function and
(c) domain score estimation.

A test developer has to adopt a particular reliability index suited to his/her purpose.

It is recommended by Hambleton et.al. (1976) that,

"If the purpose of the test is to classify examinees this mastery-non mastery states, a decision theoretic approach should be employed." Subkoviak, in his article contributed to “A Guide to Criterion-Referenced Test Construction” edited by Berk (1984) discussed the following five separate approaches for establishing the reliability of mastery-non mastery classification:

(a) Carver’s approach (1970);
(b) Swaminathan et.al. approach (1974);
(c) Huynh’s approach (1976);
(d) Subkoviak’s approach (1976);
(e) Marsall and Haertal approach (1976).

A test developer has to adopt an appropriate approach and a specific index within that approach.
2.4.11 Selection of a Standard

If the purpose of a CRT is to allocate mastery-non mastery states examinees, it becomes necessary to adopt a particular statistical method for the determination of a standard. Such a standard setting is known by various names in CRT literature. Some of these names are as follows:

Cut off score, performance standard, passing score, criterion level, threshold score and cutting score. The relevant literature on CRT contains several methods for the determination of cut-off score. These methods are classified into three categories.

(a) Judgmental;
(b) Empirical and
(c) Combination of judgmental empirical.

A test developer has to adopt a particular method form one of these three categories which serves intended purpose.

2.4.12 Test Manual

The true character of any test becomes known from the extent to which a test developer has presented the technical information about the test in the manual prepared by him/her. Moreover, a test user finds it easy to administer a test with the help of the technical information so presented. A test manual presents the true picture of a test. Therefore a good test manual should contain all relevant technical information to enable a test user to administer a test with ease and comfort. There are two types of a test manual:

(a) a test administrator's manual and
(b) technical information manual.
A test administrator's manual should contain the following information:

1. the specific administrative guidelines;
2. the oral and written direction to be given to the testees and
3. scoring procedure etc.

A technical information manual should contain the following information:

1. A technology adopted of item-writing;
2. information of item review;
3. the details about the sample,
4. the item selection procedure;
5. the information about validity and reliability procedures and
6. the information about the experts.

The next chapter presents the Review of the Related Literature of the present study.
FOOTNOTES

1 M.B.Buch (editor) : *A survey of Research in Education (First Ed.)*. Baroda : CASE, the Faculty of Education and Psychology, M.S.University, 1,1974, p-363.


25 R.K.Hambleton and et.al., op.cit, p-1112.


28 R.K. Hambleton (1982) op. cit P. 32


41 Ibid., 1975, p.958.


