CHAPTER II
LITERATURE REVIEW

Assessment and testing in the Malaysian school and college classroom for the most part has been summative. It is used to rate student progress at the end of some period of time rather than to monitor it formatively so that adjustments in instruction can be periodically made. There is little evidence to suggest in existing practice that teachers use their own tailor-made tests analytically to identify key factors upon which to build learning experiences. Hence there is a growing sense of need today to link assessment with teaching and learning, and to make instructional decisions more rational and less intuitive. Those who are skeptical tend to view testing and teaching as two mutually exclusive events which unrelated with each other (Wise, 1978). There is also the persistent notion that testing takes time away from teaching (Shanker, 1979). This results from an inability to identify the symbiotic relationship that exists between instruction and assessment. The intention of a diagnostic assessment programme in the classroom is it will lead directly to more effective classroom teaching and hence to enhanced student learning. There are arguments for a diagnostic-prescriptive model of instruction in the classroom. Goodson and Okey (1978) in a study involving undergraduate students in a physical science course found that those who were assigned to a group which was student-directed, objectives-directed with diagnostic tests and student originated discussions have the highest gains in achievement. The group that was not objectives-directed, not tested and not given remedial work has the lowest gain in achievement. His studies concluded that there was a close relationship between teaching, diagnostic testing and remedial
techniques. These three factors seemed to be a fundamental ingredient for success in using a diagnostic-prescription approach to instruction. Long (1979), who studied the biology achievement of high school students, found that initially students in the diagnostic-prescriptive groups capitalised on the opportunity to change or modify their learning deficiencies.

Black and Dockrell (1984) proposed a model for the creation of diagnostic assessment systems which involves five major components – specification of intended learning outcomes or behavioural objectives, creation of a system for diagnostic assessment, creation of appropriate diagnostic instruments in the form of criterion-referenced tests including performance standards and cut-off scores, the use of the system and instruments to obtain feedback, the use of formative and summative feedback for individual reinforcement and curriculum and teaching evaluation. The steps involved in designing and implementing the diagnostic assessment model are shown in Figure 1. This model is a variation on the basic mastery model as originally proposed by Bloom (1968).
Figure 1. A model of diagnostic assessment.

According to Black and Dockrell (1984), diagnostic assessment:

1. is a form of assessment designed primarily to help students to learn and to help teachers to teach;

2. is an application of criterion-referenced assessment;

3. is used for formative assessment purposes;
(4) provides information of student attainment of domains which are sufficiently
distinct to allow teacher to provide individual students with remedial help;
(5) is designed for all students and not solely for those with basic learning difficulties;
(6) is expected to be used as part of a system of diagnosis and consequently the
outcome of the assessment is not an end in itself but a guide to remedial action if
required.

**Instructional Objectives and Intended Learning Outcomes**

**Taxonomies of Educational Objectives**

Educators agree about the value and role of instructional objectives but
disagree about how these objectives should be stated. However, there is mutual
agreement that explicit statements consisting of specific intended learning outcomes
are more helpful than implicit statements which are general intended outcome.
Gronlund and Linn (1990) recommended that objectives should include both explicit
and implicit statements.

In the above model of diagnostic assessment, the first stage is the specification
of intended learning outcomes or objectives. This set of objectives is derived from the
statement of the content specified in the curriculum of the subject or course that is
selected for this study on diagnostic assessment. Tyler (1969) used the term
'objectives' to refer to the change in student behaviour brought about by the learning.

A taxonomy of learning objectives is a hierarchical scheme for classifying
learning objectives into various levels of complexity. It can help teachers bring to
mind the wide range of important intended learning outcomes and thinking skills. A
review of the relevant literature on taxonomies of instructional objectives reveal that
there fall into one of three domains: cognitive, affective and psychomotor. Since
assessment in colleges is mainly cognitive, the focus of this discussion is on some taxonomies for the cognitive domain.

Bloom, Engelhart, Furst, Hill and Krathwohl (1956) constructed a taxonomy of educational objectives in the cognitive domain. This cognitive domain includes those objectives which deal with the recall of knowledge and the development of intellectual abilities and skills. The taxonomy is organised into six major classes:

(1) Knowledge

This involves behaviours and test situations which emphasise recognition or recall of ideas, material or phenomenon.

(2) Comprehension

Objectives and behaviours under this category include translation, interpretation or extrapolation. Translation involves putting a communication into other terms or forms of communication; interpretation deals with reordering a communication into a new configuration in the mind of the individual and extrapolation includes estimations and predictions based on understanding of trends or conditions in a communication.

(3) Application

Application requires comprehension of the method, theory, principle or abstraction so that a student can demonstrate their uses when required to do so.

(4) Analysis

Analysis break down the material into its constituent parts, detect the relationships between parts and how they are organised.

(5) Synthesis
Synthesis involves putting together parts to form a whole. This may involve combining parts of previous learning experiences with new materials and reconstructing them into a more well-integrated whole.

(6) Evaluation

Evaluation is defined as the making of quantitative or qualitative judgments about the value or worth of ideas, methods and solutions.

Bloom et al. found that the taxonomy satisfies four criteria of usefulness.

(1) There is good agreement between education workers in the classification of specific objectives and test materials.

(2) Most statements in the cognitive domain could be placed within the classification scheme.

(3) The taxonomy stimulates thought about educational problems.

(4) The taxonomy is widely accepted by workers in the education field.

Ebel (1965) agreed that Bloom’s taxonomy is useful for classifying instructional objectives. However, he feels that using the taxonomy to classify test items will be difficult because of the unknown cognitive processes involved in answering a test question. He created a categorisation system, Ebel’s Relevance Guide, which can be used effectively for test items rather than instructional objectives. The Guide classifies test items into seven categories:

(1) Understanding of terminology or vocabulary.

(2) Understanding of fact, principle or generalisation.

(3) Ability to explain or illustrate.

(4) Ability to solve numerical problems using calculations.

(5) Ability to predict what is likely to happen under specified conditions.
(6) Ability to recommend appropriate action in specific and practical problems.

(7) Ability to make an evaluative judgment.

He further stressed that these categories classify achievement test items are useful because they are defined in terms of explicit behavioural requirements, rather than in terms of presumed cognitive hierarchical processes that may be required for successful response.

Gagne (1985) provided a taxonomy consisting of six levels of performance that need to be taught and learned. Learning objectives based on his scheme include:

(1) Discrimination

‘ability to discriminate among different stimuli’

(2) Concrete concept

‘ability to identify a stimulus as belonging to a particular class or category’

(3) Defined concept

‘ability to classify objects, events or relations and demonstrate an understanding of a concept’

(4) Rule

‘ability to demonstrate responses that indicate a rule is being applied in a variety of different situations’

(5) Higher order rule (problem solving)

‘ability of the learner to generate a new rule to solve a problem by combining two or more previously learned rules’

(6) Cognitive strategies

‘ability to adopt internal processes to choose and change ways to focus attention, learn, remember and/or think’
Nitko (1996) cautioned teachers who use a taxonomy for assessment to recognise that thinking skills may not be hierarchical and student performance on complex tasks may involve using several thinking skills at the same time. However this should not distract the use of taxonomies as tools to judge whether both higher and lower order skills of students have been assessed.

**Instructional Objectives and Assessment**

In an article on the procedure for constructing content-valid items for objectives oriented teaching and assessment, Schott, Neeb and Wieberg (1984) stressed that instructional objectives must guarantee that instructional tasks and test items are in a content-valid relationship with the instructional objectives, and that instructional tasks and test items can be constructed when instructional objectives are stated. They further asserted that the definition of instructional objectives must represent the domain specification and the behaviour specification in the following five ways:

1. The solution of the problem of content-validity depends on the degree to which the instructional objective is defined precisely. Evidence can be drawn from Hambleton (1980).

2. The instructional objectives must be formulated precisely in behavioural terms, for example, Mager (1975). Nevertheless, according to Schott, Neeb and Wieberg (1984), observability of behaviour is necessary but not a sufficient condition. Additional rules are necessary to close the gap between observed behaviour and the 'type' of action that the observed behaviour represents.

3. Instructional objectives can be realised in the processing of different items during training and testing either by using a written or oral modus. Nevertheless it must be decided whether the actual instructional element at hand corresponds to the
instructional objective. The instructional objective must be stated by means of an invariant description. The teacher must be able to determine, from the student’s response, whether he or she has reached the instructional objective. The ‘behavioural objectives’ are not suitable for this purpose.

(4) The importance of a procedure for precisely defining instructional objectives depends on their wide applicability. The smaller the group of objectives that can be defined by the procedure, the less important it will be.

(5) The last criterion is practical applicability. Teachers, without the aid of experts, should be able to apply the procedure in the classroom.

Lindvall (1967) suggested that learning objectives have to be useful for classroom instruction as well as designing assessment instruments. According to him, three minimum criteria for learning objectives need to be satisfied:

(1) Student oriented

‘Learning objectives should focus on the students. They should describe student performances’

(2) Performance oriented

‘Learning objectives should be stated in terms of what a student can perform after required learning experiences’

(3) Content-oriented

‘Learning objectives should state the specific content to which the student should apply the performance’

The flowchart shown in Figure 2 illustrates a triangle of content-validity used to compare the instructional tasks with a corresponding diagnostic test (Schott et al., 1984).
Validity

Validity refers to the ways in which test scores are interpreted and used. To validate such interpretations and uses of students’ test scores, evidence need to be provided. The *Standards* (1985) refers to three categories of validity evidence for guiding validity inquiry, namely content-related evidence, construct-related evidence and criterion-related evidence.

Content-related Evidence

Since the test items in a diagnostic model of assessment is constructed to measure a set of intended learning objectives, the content validity of the diagnostic test is of utmost importance. The *Standards* (1974) offers the following definition of content validity:

To demonstrate the content validity of a set of test scores, one must show that the behaviours demonstrated in testing constitute a
representative sample of behaviours to be exhibited in a desired performance domain. Definition of the performance domain, the user’s objectives, and the method of sampling are critical claims of content validity.

The *Standards* (1985) defines content-related validity as one that:

- demonstrates the degree to which the sample of items, tasks or questions on a test are representative of some defined universe of domain of content…. The first task for test developers is to specify adequately the universe of content that a test is intended to represent, given the proposed uses of the test.

- Another important task is to determine the degree to which the format and response properties of the sample of items or tasks in a test are representative of the universe.

Popham (1981) asserted that the focus of content validity is on the behaviours being sampled by a test. Therefore the task in demonstrating a criterion-referenced test’s content validity is to make sure that the test accurately describes the behavioural domain being measured.

According to Popham, the content validity of a criterion-referenced test can be established by a judgmental procedure involving experts in the subject matter. These experts carefully examine all items on the test and make a judgment concerning how well they sample the intended behaviour domain. For criterion-referenced tests in the classroom, the teacher is the expert.
Nitko (1996) summarises several types of evidence which can be used to judge the validity of classroom assessments. Those evidence relevant to diagnostic assessment are:

(1) Content evidence or content representativeness and relevance: How well do the assessment tasks represent the domain on important content?

(2) Substantive evidence: How much do the assessment tasks require students to use important thinking skills and processes?

(3) Reliability evidence: Will the same students obtain the same results if the assessment was applied on another occasion?

(4) Practicality evidence or efficiency and practicality: Can the assessment procedure easy for teachers to construct? Can the assessment procedure give quick results to guide instruction?

Construct-related Evidence

According to the 1985 Standards, construct-related evidence:

focuses primarily on the test score as a measure of the psychological characteristic of interest. Such characteristics are referred to as constructs because they are theoretical constructions about the nature of human behaviour.

The construct of interest for a particular test should be embedded in a conceptual framework, no matter how imperfect that framework may be. The conceptual framework specifies the meaning of the construct, distinguishes it from other constructs, and indicates how measures of the construct should relate to other variables.
The process of compiling construct-related evidence for test validity starts with test development and continues until the pattern of empirical relationships between test scores and other variables clearly indicate the meaning of the test score.

Campbell and Fiske (1959) offered a conceptual and an empirical test for construct validation. The conceptual test involves showing strong relationships between any given measure of a construct with other measures of the same construct, but weak relationships with measures of other constructs. Empirically, the strategy involves convergent validity and discriminant validity. Convergent validity refers to the correlations among different methods of measuring the same construct, while discriminant validity refers to the correlation among different constructs that are measured by the same method. For validation of a given construct, it is expected that convergent validity should be higher than discriminant validity. The different methods of measuring a given construct may be tests, questionnaires or ratings.

Messick (1988) pointed out that discriminant validity is a stronger test of construct validity than convergent validity. According to him, discriminant validity implies a challenge from an alternative hypothesis. As Messick (1975) pointed out:

'If repeated challenges from a variety of plausible rival hypotheses can be systematically discounted, then the original interpretation becomes more firmly grounded.'

He further pointed out that content-related validity is not validity in the sense that it focuses only on test forms rather than test scores, upon instruments rather than measurements. Without the test scores and measurements that is required to provide evidence for interpretation, content-related evidence does not qualify as validity.
evidence at all. However such evidence clearly do and should influence the nature of score inferences supported by other evidence.

Cronbach (1990) provides a non-exhaustive list of research techniques that might be pertinent to construct validity. These include item analyses, internal correlations, stability of scores, varying test procedures experimentally, correlation with practical criteria, correlation with other tests, and studies of group differences.

Messick (1988) pointed out that content-related inferences are inseparable from construct-related inferences. According to him, what is judged to be relevant and representative of the domain is not the surface content of test items but the knowledge, skill or other pertinent attributes measured by the items.

This is supported by Cronbach (1971) who asserted that limited content-based inferences may be made about high scorers in a test but not low scorers even though the test items are directly sampled from domain tasks. According to him, all that can be claimed is that low scorers did not perform the test successfully. There is no basis for interpreting low scores to mean incompetence or lack of ability. Further evidence need to be found to discount other plausible factors that may have caused low test performance. These include test anxiety, inattention, low achievement motivation, fatigue or low language proficiency. Hence, discriminant validity, one of the strategies for construct-validation is required.

Messick (1989b) suggested the following types of validity evidence for educational assessments and the questions that should be answered by such evidence.

(1) *Content evidence* relates to content representativeness and relevance. A good assessment should consist of tasks or test items that represent the domain of important content.
(2) *Internal structure evidence* refers to relationships among the assessment tasks or parts of the assessment.

(a) Do all the assessment tasks work together so that each task contributes positively toward assessing the quality of interest?

(b) If the different assessment tasks are supposed to provide similar information, do the results support this?

(3) *External structure evidence* refers to relationships of assessment results to the results of other variables.

(a) Are the results of the assessment task consistent with the results of other similar tasks for the student?

(b) How well does performance on this assessment task predict current or future performance on other valued tasks or measures?

(4) *Reliability evidence* over time and content domain.

(a) Will the same students obtain nearly the same scores if the assessment procedure was applied on another occasion?

(b) If a second alternate form of the assessment task with similar content were to be developed, would the students’ results be very similar?

(5) *Generalizability evidence* over different types of students under different conditions.

Will the students’ results from assessment procedure be changed drastically if they are given special incentives?

(6) *Consequential evidence*.

(a) What do we expect to happen to students if we interpret and use assessment results in a particular way?
(b) What side effects do we anticipate happening to students if we interpret and use the assessment results in a particular way?

**Criterion-related Evidence**

The *Standards* (1985) defines criterion-related evidence as that which:

- demonstrates that test scores are systematically related to one or more outcome criteria. In this context the criterion is the variable of primary interest, as is determined by a school system, the management of a firm, or clients, for example, logically, the value of a criterion-related study depends on the relevance of the criterion measure that is used.

Criterion-related evidence takes either of two forms. The concurrent evidence relates to determining present standing on a criterion measure. This evidence is useful to show that students appear in the same relative rank order on two measures. The predictive evidence relates to predicting future performance on a criterion measure. Predictive evidence is needed to show that a positive relationship exists between scores on a test (the predictor) and scores on some acceptable measures of future performance (the criterion).

The correlation between test and criterion scores is regarded as the best kind of evidence to support valid achievement-test use. This correlation provides an independent, objective validation of the subjective judgments and decisions that must be made during test development. However Ebel and Frisbie (1991) pointed out that appropriate criterion measures may be unavailable or, in many cases, appropriate criteria are difficult to measure accurately. Another problem with empirical test
validation is the uncertain validity of the criterion scores. The criterion scores need to be highly valid and reliable measures of the ability being tested. Hence, before a criterion measure is chosen, the validity of its scores has to be rigorously validated as thoroughly as the validity of the test scores in question.

**Criterion-Referenced Tests**

In his paper, Glaser (1963) emphasised the importance of interpreting test scores to provide more information about behavioural change of the individuals being tested rather than their relative performance on vague dimensions assumed to lie behind a test score. In this paper, Glaser states:

Underlying the concept of achievement measurement is the notion of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point. On this continuum as indicated by the behaviours he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement of proficiency. The standard against which a student's performance is compared when measured in this manner is the behaviour which defines each point along the achieved continuum. The term 'criterion,' when used in this way, does not necessarily refer to final end-of-course behaviour. Criterion levels can be established at any point in instruction where it is necessary to obtain information as to the adequacy of an individual’s performance. The point is that the specific behaviours implied at each level of proficiency can be identified and used to describe the specific tasks a student must be
capable of performing before he achieves one of the knowledge levels.

It is in this sense then measures of proficiency can be criterion-referenced.

Along such a continuum of attainment, a student’s score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioural repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence obtained by a particular student which is independent of reference to the performance of others.

According to Popham and Husek (1969) criterion test score information provides individual and programme decisions arising from objectives based instructional programmes. A criterion-referenced test measures the extent to which a student has mastered a criterion. This criterion may be a pre-requisite to his commencing a new training programme. Similarly, the extent to which a student has attained a set of instructional objectives in a learning module can be measured using a criterion-referenced test. Theoretically, at the end of an instructional module of instruction, it is hoped that all learners display maximum proficiency on measures reflecting on intended learning objectives. In this sense criterion-referenced test is an absolute indicator. On the other hand, norm-referenced measures ascertain an individual’s performance in relation to other individuals on the same measuring instrument. Meaningful scores arise from the comparison. An individual student is compared to
some normative group. As a diagnostic tool, norm-referenced measures lack the specificity of criterion-referenced testing in providing information on which intended learning outcomes have not been mastered by an individual. Remedial action can then be taken accordingly.

Hambleton (1978) came out with a nine-step model for developing and validating criterion-referenced tests for classroom use. The steps are as below.

(1) Preparation of domain specifications.

(2) Preparation of criterion-referenced test specifications such as selection of domains to be measured by the test and the number of test items.

(3) Writing test items that are matched to the domains of instructional objectives.

(4) Constructing test items.

(5) Determining content validity

   (a) Use of content specialists

   (b) Use of item analysis data

(6) Test assembly

   (a) Determination of number of test items/domain

   (b) Test item selection

   (c) Preparation of directions and sample questions

   (d) Layout and test booklet preparation

   (e) Preparation of scoring keys

   (f) Preparation of answer sheets

(7) Selection of cut-off scores (if instructional decisions are to be made)

(8) Ongoing collection of reliability, validity and norm information
Setting Performance Standards

In his paper, Glaser (1963) did not mention a performance standard or cut-off scores between mastery and non-mastery of behavioural objectives. He stated that there is assumed to be a 'continuum of knowledge acquisition ranging from no proficiency at all to perfect performance' and the 'degree of competence attained by a particular student' is what is assessed. However, the earlier publication of Mager (1962) included the idea of performance standard in attaining the behavioural objectives. The passage in his text most pertinent to the setting of performance standard in criterion-referenced testing is as follows:

'If we can specify at least the minimum acceptable performance for each objective, we will have a performance standard against which to test our instructional programs; we will have a means for determining whether our programs are successful in achieving our instructional intent'.

Popham and Husek (1969), in their paper on criterion-referenced testing, used Mager's term 'performance standard: criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., performance standard.

Similarly, the notion of performance standard was included in the definition of 'criterion-referencing' by Glaser and Nitko (1971):

'A criterion-referenced test is one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards. The performance standards are usually specified by defining some domain of tasks that the students
should perform. Representative samples of tasks from the domain are organized into a test. Measurements are taken and are used to make a statement about the performance of each individual relative to the domain.

Hambleton et al. (1978) distinguished two purposes of criterion-referenced testing: estimation of 'domain scores' and 'allocation of examinees to mastery states.' The former purpose is to determine the percentage of test items in some well-defined domain that an examinee can answer. The second purpose involves the setting of cut-off scores to determine the mastery level.

Criterion-referenced tests and cut-off scores are used in instructional programmes that are based on objectives. Objectives-based programmes are designed to improve the quality of instruction by defining the curriculum in terms of objectives, relating instruction and assessment closely to objectives, providing for formative evaluation and remedial work. Evidence suggested that full and proper implementation of objectives-based programmes have resulted in greater student achievement than more traditionally –oriented curricula (Torshen, 1977).

Since decisions are to be made based upon the descriptive information provided by criterion-referenced tests, it is necessary to find out whether the performance level of a student indicate whether he or she needs remedial work on the objective defined by the teachers of the curriculum. In order to do this, it is necessary to set standards or cut-off scores. Carefully developed cut-off scores can contribute substantially to the success of diagnostic assessment in the classroom. These scores provide a basis for effective instructional decision-making.
Burton (1978) identified three methods of setting performance standards for criterion-referenced tests. They are theory-based standards, standards based on expert consensus, and standards based on practical necessities.

**Standards Based on Theories**

The original conception of criterion-referenced testing by Glaser (1963) was closely related to learning hierarchies.

Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual’s achievement level falls at some point on this continuum as indicated by the behaviours he displays during testing.

Burton (1978) interprets Glaser’s learning hierarchies requires each step or set of steps is a necessary and sufficient condition of the following steps. If test scores could be mapped onto their hierarchy of behaviours, then each final score would imply that the respondent can do the corresponding task and all tasks below it in the hierarchy. She opposed such a hierarchy of learning on the grounds that most models for learning do not posit such a simple, linear, step-by-step method of learning. Furthermore, it is highly impractical to implement such a learning hierarchy.

**Standards Based on Expert Consensus**

Shepard (1976a) suggested setting standards based on the personal experience of judges is not justified and may be biased. Justification and consequences of the performance standard must be established by some external evidence. He advocated that judges should be given all the evidence they can use. According to Shepard (1976b):
Instead of relying on their experience, which may have been with unusual students or professionals, experts ought to have access to representative norms....Of course, the norms are not automatically the standards. Experts still have to decide what 'ought' to be; but they establish more reasonable expectations if they know what current performance is than if they deliberate in a vacuum.

Burton (1976) argued that Shepard's argument will imply that the classroom teacher is the most qualified to set performance standards, since he or she has more complete information compared to external judges. She further stressed that when expert opinion is necessary, the most appropriate expert is the classroom teacher who has access to most of the relevant information.

Popham (1978) supports the position that reliance on human judgment as the main ingredient of standard-setting should be looked at positively, not negatively. According to him, serious standard-setting efforts represent genuine attempts to do a good job in deciding what kinds of standards we ought to employ. Teachers, based on their experience with a curriculum, using high quality criterion-referenced tests, should be qualified to set cut-off scores for the monitoring of student progress in their courses. Popham also pointed out that students who are misclassified into a non-mastery category due to performance below the cut-off score could not be so far above the mastery cut-off score. On the other hand, those students who are misclassified because they scored above a cut-off score, they will be tested again. It is highly likely that the error will be discovered the next time, particularly if the objectives are sequential. Ebel (1978) stressed on the importance of valid and reliable tests to reduce errors of measurement. He commented that:
Pass-fail decisions on a person's achievement in learning trouble some measurement specialists a great deal. They know about errors or measurement. They know that some who barely pass do so only with the help of errors of measurements. They know that some who fail do so only with the hindrance of errors of measurement. For these, passing or failing does not depend on achievement at all. It depends only on luck. That seems unfair, and indeed it is. But, as any measurement specialist can explain, it is also entirely unavoidable. Make a better test and we reduce the number who will pass or failed by error. But the number can never be reduced to zero.

Block (1978) expressed the view that although techniques of setting cut-off scores may be imperfect, they are not useless. He asserted that the setting and use of cut-off scores can have a decided positive impact on student learning. In their review of relevant literature on mastery learning strategies, Block and Burn (1976) found that group-based, teacher-paced strategies make extensive use of cut-off scores to monitor how a student's learning is changing over a sequence of instruction and to modify the instruction should the student not be progressing as expected. Essentially, at the end of one- to two-week segments of instruction, each student is administered an objectives-based, diagnostic progress test upon which some cut-off score has been set to indicate mastery. If the student's test score is equal to or above this cut-off, the student is judged to be progressing satisfactorily. If not, the student is judged to be progressing unsatisfactorily and is given remedial instruction for bringing his or her progress to the standard expected. Their review also indicated that students who learned for mastery usually learned more than students who did not learn. More
importantly, the experimental research suggested that the setting of cut-off scores lay at the heart of the many classroom successes of the mastery strategies. These studies indicated that the use of particularly high standards was preferable to the use of none at all in terms of student cognitive development.

**Standards Based on Practical Necessities**

A cut-off score is a point on a test score scale that is used to sort examinees into two categories, masters and non-masters, that distinguish different levels of achievement relative to a particular objective measured by a test. Glass (1978) and Burton (1978) referred to the use of standards or cut-off scores as interpreting examinee performance on each of the objectives measured by a test. Hambleton (1978) asserted that if the cut-off scores set are appropriate for some particular use of criterion-referenced test, teachers can be trained to use cut-off scores properly.

In a controlled research done by Block (1970) it was noted that a mastery level of about 80% - 90% is substantially better than a level that is higher or lower. A mastery level that is set below 80% may enable students who do not have sufficient mastery of the learning objectives to proceed effectively with instruction. Results of his study suggested that setting mastery levels high at 95% may be best for cognitive learning but, in the long term, student's interest is likely to diminish. He also found that setting mastery level at 85% may reduce cognitive learning but selected affective outcomes were maximised.

The setting of mastery at 80% for a criterion-referenced test is supported by Black and Dockrell (1980). They justified this cut-off score based on the argument of 20% reliability of test items in a test. The teacher setting a criterion-referenced test mastery score at 80% is fairly certain that those who attain mastery have, in fact,
attained the unit objectives, since the lowest estimate of their true scores is 60% and the highest estimate is 100%. However, they stressed that there is a margin of error involved. A test score of 70% may represent a true score of 50%, 60%, 70%, 80% or 90%, of which all but the first is a potential mastery when the mastery level is set at 80%. This therefore introduces a margin of error. However they argued that in the diagnostic context, a student who has mastered the intended learning objectives but appears to fail bears the consequences of extra remedial procedure. Provided that the remedial work allows a student who was wrongly classified to move rapidly back into the unit of instruction, the potential negative effect of reducing motivation will be kept to a minimum.

Angoff (1971) proposed a method for determining the cut-score that requires judges to estimate the probability that a minimally competent examinee will respond correctly to each test item. By summing these predicted responses, a predicted test score can be obtained for each judge. The mean of these predicted scores is used as an estimate of the cut-score. In a study, quoted in Sizmur (1997), to determine the cut-score for a test of Level 2 reading comprehension developed by the School Curriculum and Assessment Authority (SCAA) in the United Kingdom it was found that the 13 judges' ratings varied between 1.2 marks and 6.7 marks with a standard error of 0.45. This suggested a reasonably reliable outcome. However it was considerably lower than the cut-score of 9 marks that was finally adopted. In a field test, quoted in Mills (1983), to determine the cut-score for a second grade basic skills test, the Louisiana Department of Education in the United States found that although the mean cut-score for the group of judges were similar, their ratings were not homogeneous. There was a wide range on the individual standards on every test that
was used. According to Mills (1983), the Angoff method depends on the judges’ understanding of the meaning of minimal competence and their understanding of how a minimal competent student would respond to the test items. This may be too demanding on the judges and lead to inconsistency in judgement on the cut-score.

**Discrimination Indices for Item Analysis**

In classical test theory, the parameters utilised for selecting items for the final version of a test are based on two item statistics: difficulty level and discrimination level. High discrimination is interpreted as a desirable characteristic of an item and a key indicator of item quality. Ebel (1979) emphasises the importance of selecting highly discriminating items for inclusion in an improved version of a test. A highly discriminating item will effectively differentiate between the performance of individuals of higher and lower ability on the item.

Glaser and Nitkon (1971) define a criterion-referenced measurement as one that is deliberately constructed so as to yield measurements that are directly interpretable in terms of specified performance standards. The interpretation of performance is based on an *a priori* criterion. In a teacher-made test, this criterion refers to a domain of instructional objectives that need to be mastered. A norm-referenced measurement, unlike the criterion-referenced measurement, is interpreted in terms of the relative performance of students. Therefore, item discrimination indices that are used to evaluate test items in a norm-referenced test need to be modified.

Popham and Husek (1969) point out some limitations of norm-referenced discrimination indices if they are used for the purpose of evaluating test items of a criterion-referenced test. In a norm-referenced test, a high item discrimination is
interpreted as a desirable characteristic of an item and a good indicator of item quality. The higher the item discrimination index, the bigger the score variability and the better the item is in discriminating or differentiating between students of higher and lower ability (i.e., the greater the differences in the performance of students of higher and lower ability on the item). However, a good valid and reliable criterion-referenced test administered after adequate instruction may result in very low score variation. This discrepancy suggests that norm-referenced item discrimination indices are unsuitable for criterion-referenced tests. A typical norm-referenced item discrimination index correlates item score to total test score. An item with a low index is suspect because it has failed to discriminate between the weak and the better students. However, Popham and Husek (1969) question such an interpretation as inappropriate for a criterion-referenced test. In a criterion-referenced measurement, if an item accurately measures an essential instructional objective and all students answered correctly or incorrectly, the item cannot be rejected.

In general there are two classes of discrimination indices: upper-lower discrimination indices and correlational discrimination indices. Upper-lower discrimination indices informs on the number of discriminations made by an item, while correlational discrimination indices measures the relationship between item scores on some criterion variable, usually total test score. In the opinion of Brennan (1972), the upper-lower discrimination indices are preferred on the basis of easier interpretability and computation.

**Brennan Discrimination Index**

Brennan (1972) constructed a discrimination index that differs from that proposed by Findley (1956). Brennan's index does not require equal number of
students in upper and lower groups. Moreover the requirement that the test scores be
normally distributed is unnecessary. According to Popham and Husek (1969), in a
criterion-referenced test, the test constructor expects most of the students to get most
of the items correct, yielding a distribution of test scores that is negatively skewed.
According to Brennan (1972) the discrimination index $B$ has the advantage of ease of
interpretation as compared to correlational indices. In terms of interpretability, a $B$
value of $+1$ is twice as large as a $B$ value of $0.50$. On the other hand, a correlation
coefficient $r$ of $+1$ does not indicate a discrimination that is twice as large as a $r$
of $0.50$.

**Point-biserial Correlation, $r_{pb}$**

The point-biserial correlation coefficient is used as a measure of item
discrimination for items that are scored dichotomously, right or wrong. The
dichotomous right or wrong response to a test item is quantified with arbitrarily
applied numerical values 1 and 0 respectively. The test scores of students are
continuous. Their scores are correlated with their response to each item on a test.

The value of the point-biserial correlation coefficient ranges from $-1$ to $+1$.
A $r_{pb}$ of $+1$ indicates that those who score highest on the test tend to get the item
correct and those who scored lowest tend to get the item wrong. Similarly a $r_{pb}$ of $-1$
indicates that those who scored lowest tend to get the item right and those who get the
item wrong are students who scored highest. A $r_{pb}$ of 0 means that there is no
correlation between scores on item and scores on test.
**Phi Coefficient, $\phi$**

If an item which is dichotomously scored, right or wrong varies with another dichotomy of mastery and non-mastery, the correlation between the two variables may be determined by using the *phi coefficient, $\phi$*.

An item is considered poor if there is little difference between the numbers of masters and non-masters who passed or failed the item. Swezey and Pearlstein (1975), quoted in Swezey (1981) suggested that if the $\phi$ value of an item is less than +.30, or is negative, the item may be a poor one. According to them a $\phi$ value of less than .30 indicates that the item does not discriminate well between masters and non-masters. A negative $\phi$ value means that non-masters generally did better on the item than masters did. Black and Dockrell (1984) studied the use of the phi coefficient to classify an individual as master or non-master based on the observed scores of a diagnostic test. They found that although the coefficient gave a rough indication of problem items, it appears to lose valuable information which could be gained from an analysis of observed data. They suggested that developing ground rules for making decisions in relation to observed raw data was more effective in obtaining information of items that may gave create problems in the test.

**Interpretation of Discrimination Indices with Criterion-Referenced Tests**

For norm-referenced tests, non-discriminating items and negatively discriminating items are unacceptable while positively discriminating items are acceptable. Obviously an item that does not discriminate tells the teacher nothing about differences among students. Negatively discriminating items are flawed since the weak students get these items right while the best students seem to get them
wrong. A positively discriminating item is acceptable since it identifies student differences correctly.

Popham and Husek (1969) suggest that the interpretation of discrimination indices needs to be modified for criterion-referenced tests. If an item tests on an important objective, this item is an effective item even though the item does not discriminate between poor and good students. Negatively discriminating items are interpreted in the same way as in norm-referenced tests. They are flaws in such items. However, Popham and Husek (1969) stressed that although a positively discriminating item is acceptable, it may suggest some weaknesses in the instructional strategy. According to them if the instructional strategy is equally effective for all students, the discrimination index should be zero and not positive. Thus, a positive discrimination index may indicate a need for revision of teaching strategy in order to be more effective for the weaker group. On the other hand, if teacher instruction is equally effective for all students, a positive discrimination index may indicate the need for item revision. Hence, they argued that a positive discrimination index is not ideal because it may indicate weakness in instruction or in the item.

**Mastery Learning**

The diagnostic assessment model is closest to the approach used in mastery learning. The idea of developing instructional programmes to provide successful learning experiences for students has resulted in the development of a diverse collection of models. These models offer new approaches in learning that are aimed at maximising the opportunity to learn new materials in the school or college curriculum. Among the widely implemented models is Mastery Learning (Block, 1971; Bloom, 1968; Caroll, 1970). Many studies have been done since Bloom (1968)
presented his paper on mastery learning. The findings suggested that the mastery model can be economically and easily implemented in courses at any level of education and in a wide range of content area (Block, 1970).

Mayo (1970), in describing the mastery learning model, notes that:

(1) Students are made aware of course and unit expectations, so that they view learning as a cooperative rather than as a competitive venture.

(2) Standards of mastery are set in advance for the students, and grading is in terms of absolute performance rather than relative performance.

(3) Short diagnostic tests are used at the end of each instructional unit.

(4) Additional learning is prescribed for those who do not demonstrate unit mastery.

(5) Additional time for learning is prescribed to students who seem to need it.

An important component of the mastery learning paradigm is diagnostic testing. Diagnostic testing is an essential part of the teaching-learning process in the classroom. It is designed for all students and not solely for those with basic learning difficulties.

Block (1971) notes that to individualise instruction within the context of ordinary group-based instruction, mastery learning relies heavily on the constant flow of feedback information to teacher and learner. A flowchart of steps used in monitoring student progress in a typical version of a mastery learning programme (Hambleton, 1974) is shown in Figure 3

In the flowchart, the unit posttest is a formative test. According to Scriven (1968), such a test is used exclusively for diagnosing learning difficulties faced by students in the unit of instruction. A formative test, or alternately called a diagnostic test is a criterion-referenced test that is designed to cover the intended learning
objectives over a unit of instruction in the classroom. It is used to determine whether or not a student has mastered the material and to serves as a basis for prescribing supplemental work in area where the student is weak (Airasian, 1971).

Implicit in educational psychology is the assumptions that individuals differ widely in their intellectual abilities. These differences are reflected in the outcomes of schooling. However, advocates of mastery learning such as Bloom (1976) and Block (1974b) have challenged these assumptions. They claimed that individual differences in learning ability are unreal and that considerable greater degrees of learning equality are possible. They argued that traditional instruction holds time constant and allow achievement to vary. However in mastery learning, time is allowed to vary and variation in student achievement level is reduced. Ultimately mastery learning aims to reduce or eliminate any correlation between aptitude and achievement. However two problems may arise. Firstly, if non-masters take much longer to achieve a particular criterion level then corrective instruction must be given outside of regular class time. Secondly, masters have to spend large amounts of time waiting for non-masters to catch up.
Mastery learning is an instructional technique for the teaching and learning of hierarchical sequential material. The learning module is subdivided into smaller units of instruction. Students are given a test at the end of each unit, and if they do not achieve a mastery grade on the test, they are provided with remedial teaching until they can achieve a mastery grade on a retest. In a review of 157 mastery learning
studies in which the learning outcome was cognitive achievement, he found that results of 107 of the studies were statistically significant in favour of mastery learning while only 3 studies favour non-mastery approaches. Similarly, Hyman and Cohen (1979) found that mastery learning was consistently more effective than traditional approaches to learning. Bloom (1978) argues that individual differences in learning rates and time needed to learn are the outcomes of the schooling process. He claims that students lacking in the prerequisite knowledge, will have a slower rate of progress in a new learning unit. Bloom (1976) asserted that the rate of learning or amount of time needed to learn some criterion of achievement is an alterable characteristic. His studies suggested that the differences between good and poor learners, or fast and slow learners, could be reduced. Block (1974b) claims that, under appropriate instructional conditions, individual differences in rate of learning are very small.

In a met-analysis of findings from 108 controlled evaluations, Kulik, Kulik and Bangert-Drowns (1990) showed that mastery learning programmes have positive effects on the examination performance of students in colleges, high schools and the upper grades in elementary schools. On the average it was found that mastery learning programmes raise final examination scores by about 0.5 standard deviations in these colleges and schools. The effects appear to be stronger on the weaker students in a class. They also found out that mastery programmes have positive effects on student attitudes toward course content and instruction as compared to that of students in conventional classes. However, students spent more time on instructional tasks.

Resnick (1977) argues that mastery learning could have serious consequences for current group-based approaches to schooling. Since students proceed as groups to the next unit, faster students are likely to be held back waiting for slower students to
master each unit. Mueller (1976) claims that a disproportionate amount of resources would be funneled to slower students in the form of corrective efforts. Given the time constraint faced by the teacher, extra teacher time given to the slower students would be taken from faster students.

**Feedback**

Diagnostic tests are formative tests. Test results of students provide valuable feedback to the teacher. The teacher uses this feedback to monitor or guide student learning while it is still in progress. Nitko (1996) gave two purposes of classroom diagnostic assessment of learning difficulties:

1. To identify which learning objectives a student has mastered or has not mastered, as well as to decide how instruction needs to be adapted to the students.

2. To suggest possible causes or reasons why the class as a whole or an individual student has not mastered the learning objectives, and decide what might need to be reinforced.

He further presented six approaches to diagnosing learning difficulties:

1. Profiling the content areas strengths and weaknesses of a student who has low achievement standing relative to his peers.

2. The lack of prerequisite knowledge or skills deficit approach.

3. The mastery of specific learning objectives approach.

4. Identifying students’ errors in performance approach.

5. The knowledge structure approach to identify deficit in learning due to a student’s inappropriate cognitive organisation of concepts.
(6) The component competencies of problem solving approach, in which a learning
deficit is the student’s inability to perform one or more of the components
necessary to answer a test item.

Although feedback is an important component of the teaching-learning process,
there are situations in which feedback might have no benefit for learning (Snow,
1972). When students are capable and willing to produce their own feedback or when
a task is so easy that performance feedback is unnecessary, feedback from external
source may have a negative impact on learning. This may be cognitively and
motivationally inhibiting (Corno & Snow, 1986). According to Anderson, Kulhavy
and Andre (1972), learning is promoted if feedback is given only after students have
generated responses to given questions and problems. This is supported by Salomon
and Globerson (1987) who found out that feedback can promote learning if it is
received mindfully, but it can inhibit learning if it encourages mindlessness. On the
basis of his review of research on feedback effects, Kulhavy (1989) theorises that a
student’s response to test-like items tends to naturally persist. Hence, feedback’s most
important instructional effect would be to correct erroneous responses not to
strengthen correct responses, as the behaviourists argued.

In a meta-analytical review of 53 comparisons of instructional treatments with
immediate or delayed feedback, Kulik and Kulik (1988) found that in tasks that
involve complex conceptual learning, immediate feedback was superior to delayed
feedback. Bangert-Drowns, Kulik, Kulik, and Morgan (1991) did a meta-analysis of
forty studies which yielded 58 feedback effect sizes. In these studies each effect size
represents a standardised comparison between an experimental group that received an
instructional treatment with feedback and a control group that received identical
instruction without feedback. They found that the type of feedback was strongly related to effect size. When feedback merely indicated that a response was correct or incorrect, it resulted in a lower effect than when the feedback in some way informed the learner of the correct answer. The studies also showed a positive correlation between error rate and effect size. As students make more errors during instruction, feedback has a more important effect on later retrieval of correct information.

Diagnostic assessment in the classroom involves frequent testing as compared to the conventional method. These tests are used during instruction to evaluate and guide student progress. Students who do not show mastery on such tests have their weaknesses diagnosed, received remedial instruction and are given new opportunities to show mastery. Ordinary classroom tests are often used without feedback and correctives as extensive as that used with diagnostic testing or mastery testing. With ordinary classroom tests, test performance is a one-time event that contributes to the student's academic record. In a meta-analysis focused on research on frequent classroom testing, Bangert-Drowns, Lin, Kulik and Kulik (1991) showed that the use of classroom testing does increase performance on criterion measures of achievement, but at a diminishing rate of returns. When tested groups were compared with groups who received no tests, the tested group typically scored about one half standard deviation higher on a criterion examination than did the untested students. It was also found that when two groups of students answered identical test items, students performed better if they answered the questions on a large number of short tests rather than on a small number of short tests. However these studies cannot draw conclusions about effects of frequent testing because they do not investigate performance on a common criterion examination given to both experimental and control groups under
the same condition. Finally, this meta-analysis shows that teachers can improve the affective outcomes of instruction by testing students more often. Frequently tested students were found to have a more favourable opinion of their instruction. Increasing the frequency of tests may be a way of creating a more positive atmosphere in the classroom.

Butler and Winne (1995) proposed five functions that feedback serve. First, when students' conceptual understandings or beliefs are consistent with instructional objectives, feedback can confirm that condition. Second, if students lack information, feedback can help students add information, thereby elaborating and enriching prior knowledge. Third, where elements of prior knowledge are incorrect, feedback can provide information to replace those propositions. Fourth, if students' understanding are basically correct, they still may need to tune those understandings, for example, by discriminating between concepts. Fifth, if students hold false theories that are incompatible with new materials to be learned, they may need to completely restructure schemata with which information in the domain is represented.

A Diagnostic Assessment Model

Based on a review of the relevant literature, a diagnostic assessment model as shown in Figure 4 is developed.

Description of the Model

The model moves from top to bottom identifying paths between instruction, assessment, item analysis and feedback for decision-making. The model represents diagnostic assessment as a systematic process and objective-oriented. The model also represents diagnostic assessment as dynamic. The instructor reflects on the test and item analysis data before making mastery and non-mastery decisions on achievement.
Implementation of the Model

The steps in the implementation of the model are as follow:

(a) The model recognises that learning objectives is an important ingredient of this model, including planning what objectives to emphasise in instruction as well as to how to teach these objectives in the classroom. Before the implementation of the model, the physics course content for the 19-week second semester is organised into four modules and 7 units. Instructional objectives are constructed for each unit based on Bloom's taxonomy. Instruction in the classroom focuses on student learning of objectives and includes lectures, class problem-solving activities and home assignments. Each instructional period lasts for one hour.

(b) The model views teaching and learning as a dynamic process. It recognises the importance of obtaining feedback regarding the effectiveness of the instructor's teaching practice and strategy as well as student learning of objectives and response to instruction. An essay-type or multiple-choice criterion-referenced test is given on completion of instruction for each learning unit which typically lasts for two to three and a half weeks. The time allowed for each test is 30-60 minutes. Based on a set criterion of 80 %, the students are classified as masters and non-masters. The item difficulty index and discrimination indices \( B, r_{pb} \) and \( \phi \) are computed for each item. \( B \) is used because it fairly depicts the discriminating ability of an item as well as its ease of interpretation. The learning deficiencies of the weaker students that are identified by item analysis is monitored and documented. This data is then used to devise the remedial activities and assignments to improve and enhance student's skills in areas of the course content when performance does not meet the standards set.
Figure 4. A model of diagnostic assessment