CHAPTER V
DISCUSSIONS AND RECOMMENDATIONS

Introduction

Diagnostic assessment is a practical means of ensuring that teachers adhere to a particular curriculum and students learn a specific set of objectives. Although a well-defined set of minimum learning objectives for each unit may place a restriction on the maximum learning that students might have achieved, these objectives are deemed to be essential to the course of study based on the curriculum and also the pedagogical content knowledge of the teacher. The diagnostic assessment model implemented in this study contains certain similarities to the mastery learning model such as setting instructional objectives, administering criterion-referenced tests and mastery and non-mastery classification of students. One of the differences between the two models is that diagnostic assessment includes a post hoc item analysis of test data using difficulty and discrimination indices to make more specific monitoring decisions, attainment decisions and diagnostic decisions. to improve student learning and teaching effectiveness. The other difference is that, unlike in group mastery learning where the teacher could not move ahead until virtually all students have demonstrated mastery of the prerequisites, diagnostic assessment has the more modest aim of using the post hoc test item analysis data to remedy learning difficulties, give group corrective instruction and modify teaching strategies before advancing to the next unit of instruction. Diagnostic assessment therefore represents an effective means of providing quality group corrective instruction or remediation without bearing the cost of a slower instructional pace as in mastery learning. It
overcomes one major problem of mastery learning i.e., the problem of faster students being held back by sitting through unnecessary class remedial instruction time and waiting for slower students to master their individualised remediation. Since diagnostic assessment possess some features of mastery learning while differing in others, the findings of this study is compared to the findings of past studies on mastery learning to see whether there is any agreement between the two models in improving learning outcomes.

**Academic Achievement**

**Comparison between groups.** The mean posttest score for the low achievement group is lower than that of the medium and high achievement groups even after the treatment of diagnostic assessment. One possible explanation for this findings is that the amount of corrective individual assignment and instruction on concepts may not be enough to remediate the learning deficits of low achievers. The corrective feedback is given in group settings. This may be enough to increase the achievement scores but not enough to identify and remedy serious deficits. Studies of students' pace through individualised materials routinely find that the slowest students require 200 - 600% more time than the fastest students to complete the same amount of material (Arlin & Westbury, 1976). This is far more than what diagnostic assessment is likely to be able to provide in this study. Another reason could be that the feedback-corrective cycle is simply insufficient in itself to produce a substantial improvement in the achievement of the low achievers. Bloom (1984) noted that there are many variables other than feedback-correction that influence an effective instructional programme. These variables, instructional quality, adaptation to
individual needs, motivation and instructional time, may need to be considered to produce a large effect on academic achievement. (Slavin, 1987)

The findings of this study showed that diagnostic assessment appears to have very positive effects only for the low achievers. However the study found that the implementation of the model in the classroom is not detrimental to the achievement of medium and high achievers. Therefore diagnostic assessment, which incorporates certain elements of mastery learning - in particular, assessments of student learning of specific instructional objectives and corrective feedback, allows for immediate remediation of specific learning deficits of low achievers before they accumulate into large deficiencies. This may have led to the significant improvement in their scores.

**Variation in examination performance.** The pretest-posttest scores indicated that there is a decrease in the standard deviation from 26.51 to 20.83, a reduction of 21%. The variation of posttest scores was 79% of the variation of the pretest scores for the group. This reduction with student variation in examination performance scores agrees with past research on testing for mastery. Bangert-Drowns, Kulik and Kulik (1990), found that 37 out of 52 studies reported a decrease in the variation of examination scores as a result of learning for mastery. The average variation of examination scores of the mastery learning group was 77% of the variation in scores of the control group.

**Effect size.** In a meta-analysis of 36 studies that used Bloom's learning for mastery approach to learning, Bangert-Drowns, Kulik and Kulik (1990), found 34 studies that produce results that were favourable to the experimental treatment. They found that the average effect of learning for mastery was to raise examination scores
by 0.59 standard deviations i.e. an effect size of 0.59. In their analysis, effect size is defined as the difference between the mean scores of the experimental and control groups divided by the standard deviation of the control group. In the present study, the experimental group of students serves as the control group itself. The effect size of the diagnostic assessment used in this study is calculated from the $t$-values obtained in the correlated $t$-tests. The effect size of diagnostic assessment on the low achievement group is 1.02. This increase in the student achievement scores of 1.02 standard deviations strongly supported the implementation of diagnostic assessment in the classroom since it appears to have a positive effect on learning for the low achievers in the class. Surprisingly, there is a negative effect size of -0.64 for the high achievers. The findings suggested that effects of diagnostic assessment are not uniform on all students in a class. Low aptitude students appear to have benefited from the assessment whereas high aptitude students showed a deterioration in their performance. The decrease in performance of the high achievers may be attributed to statistical regression towards the mean as shown by the pattern of test scores before and after the implementation of diagnostic assessment.

The findings of this study for the low achievement group is supported by past studies using the learning for mastery approach. Bangert-Drowns, Kulik and Kulik (1990), in a meta-analysis of 108 studies, found that the average improvement in scores of low aptitude students is 0.61 standard deviations. The findings of this study for the high achievement group however do not concur with their findings which showed an average improvement in scores of 0.40 standard deviations for the high aptitude students.
Frequency of testing. Although diagnostic assessment is an effective means of holding teachers and students to a valuable set of instructional objectives, there is still the question of which elements of diagnostic assessment account for its effects on performance. There is some evidence that much of this effect may be accounted for by frequent testing and feedback to students rather than the entire feedback-remediation cycle. Kulik, Kulik and Bangert-Drowns (1990) report that mastery learning studies that failed to control for frequency of testing produced mean effect sizes almost twice those associated with studies in which mastery learning and control classes were tested with equal frequency. Long, Okey and Yeany (1978) compared mastery learning to a condition with the same frequency of testing and found a smaller effect than in a comparison with a control group that did not receive tests. Hence, the higher frequency of testing in diagnostic assessment may have contributed to the improvement in performance of the students. Ordinary classroom tests are often given less frequently. They are also used without feedback and corrective instruction as extensive and specific as that used in the present model which uses item-analysis as an effective tool to pin-point the main weaknesses of student learning of objectives. This is supported by the findings of past studies. In a meta-analysis of findings from 35 studies on frequency of classroom testing using criterion test performance, Bangert-Drowns, Kulik and Kulik (1991) found that 29 of these studies resulted in improvement of performance with an average effect size of 0.23. Thirteen of these 29 studies reported a significant difference in the posttest performance between experimental and control groups. Only one of the negative findings was statistically significant. These results indicated that frequent classroom testing is beneficial to student achievement. The effect sizes were almost invariably high when the
frequently tested group was compared with a control group that received no tests. In 11 such studies, average effect size was 0.54. In eight studies that compares the high, intermediate and low frequencies of testing, it was found that the high frequency group scored higher on criterion examinations than did the intermediate frequency group in seven of the studies. The average effect sizes for the high frequency group with 23 tests was significantly higher than that of the medium frequency group with 7 tests (0.49 > 0.23). This findings suggest that the use of classroom testing does increase performance on criterion tests, but at a diminishing rate of return.

In this study, there were four tests given before the first semester examination (pretest) and seven tests before the second semester examination (posttest). The pre-treatment tests are summative, longer and cover more than one topic per test whereas the post-treatment diagnostic tests are formative, shorter and cover only one unit of instruction per test. This appears to have contributed to the improvement in performance as a result of the treatment. Bangert-Drowns, Kulik, Kulik and Morgan (1991) found that when two groups answered identical test items, superior performance was obtained from students who answered the questions on a large number of short tests rather than on a small number of long tests. However this study cannot draw the same conclusion as above because the tests for the first semester and second semester do not investigate performance on the same set of instructional objectives.

Prior knowledge. The implementation of the diagnostic assessment model in this study involves organising the physics syllabus into units of instruction. These units are sequential in the sense that mastery of one unit does provide the prior knowledge for a student to follow the next unit of instruction effectively. Ensuring
that students have the necessary prior knowledge before commencing instruction may have been a contributing factor to the improvement in student performance as a result of diagnostic assessment. The effect of prior knowledge on learning outcomes has been shown in past studies that try to explain the variance in test scores. Knowledge measured prior to a course, explained, on the average no less than 50% of the variance in the posttest scores (Dochy, Moerkerke & Martens, 1996). Bloom (1976) reported correlations of 0.50 to 0.90 between pretest and posttest scores. Dochy (1992) found that in real life classroom settings prior knowledge accounted for up to 42% of the variance in posttest scores. Weinert (1989), quoted in Dochy, Moerkerke and Martens (1996), found that domain specific knowledge is a decisive prerequisite for good mathematics achievement. Walker (1987), quoted in Dochy, Moerkerke and Martens (1996), reported that domain specific knowledge can compensate for low intellectual ability, but a high intellectual ability cannot compensate for a low prior knowledge. Therefore diagnostic assessment, which measures the prior knowledge state of students using criterion-referenced tests, serves as a tool to build up their prior knowledge required for the following unit or units of instruction.

**Performance Standards**

In this study, a passing percentage test score of 80% is used operationally to make mastery/non-mastery decisions about students. It is assumed that there is no distinction between passing score and performance standard, which is a conceptual boundary between acceptable and unacceptable levels of achievement. However, passing-score is merely a number while performance standard is a construct. The passing score for a unit test must be chosen to represent some intended distinction between masters and non-masters of the content domain of a unit of instruction. Non-
masters are judged to be unprepared to function effectively because they lack mastery of some content and skill in performing certain tasks. Glass (1978) has commented on the arbitrariness of performance standards, while others have suggested that, although passing scores are arbitrary in the sense that they are based on judgment, they are not capricious (Popham, 1978; Block, 1978; Hambleton, 1978). The score scale in a test is generally a continuous function of level of achievement. As a result, there is no simple way to choose a particular point on the score scale as the passing score. So there is some arbitrariness in setting the passing percentage score at 80% for all the tests given during the treatment of diagnostic assessment. As noted by Jaeger (1990), quoted in Kane (1994), students just below the passing score do not differ substantially from students just above the passing score. Therefore students who achieve a score at or above 80%, may or may not have met the corresponding performance standard of mastery of the instructional objectives for a unit. If the correspondence between passing scores and performance standards is excellent, then it is to be expected that students who are classified as masters in the diagnostic tests should perform equally well in the final summative examination. The correlations between the tests and the examination range from .41 to .83, giving a mean value of .66 (Refer Table 12). This high correlation shows that the choice of a criterion score at 80% to be associated with the performance standards expected of students is both practical and reasonable. It also indicated that the diagnostic tests possess good predictive validity.

**Attitudes and Perceptions**

The above achievement effects of diagnostic assessment on the low achievement group are accompanied by affective and motivational gains as indicated
by their ratings on the two attitude scales - Semantic differential scale and Likert-scale. The findings of the Likert-scale showed a significant difference in the perceptions of the low achievement group and the high achievement group. The low achievement students perceive diagnostic assessment most favourably while the high group has the least favourable perceptions. Overall, the groups perceive the treatment positively. This is supported by the responses of these two groups of students on the evaluation dimension of the semantic differential scale. Diagnostic assessment was perceived positively by both groups. However, the low achievement students indicated a stronger positive feeling towards diagnostic assessment as compared to that of the high achievement students. These findings concur with previous studies on the effects of mastery learning programmes on student attitudes toward the subject matter that were being taught. In a meta-analysis of 36 studies on the effectiveness of mastery learning programmes, Bangert-Drowns, Kulik and Kulik (1990) found 12 out of 14 studies reported that student attitudes were more positive in mastery classes than in conventional classes. The positive correlation between achievement gains and affective gains may be due to the fact that affect interacts with student learning strategies. Lai and Biggs (1994) suggested that when students are provided with instructional support sufficient to require their mastery of objectives they feel better about school, teacher, subject and self. These students will also be better prepared cognitively and emotionally for subsequent learning tasks. (Bloom, 1981).

**Intrinsic motivation.** Lens and Decruyenaere (1991) distinguished three categories of intrinsic motivation for learning. The category that is of interest here has to do with man as a problem solver, looking for challenge. Preserving and increasing self esteem is central to the notion of challenge (Dweck, 1986). Based on this theory
of intrinsic motivation, learning tasks should be presented by the teachers and perceived by students as achievement tasks. In diagnostic assessment, mastery of intended outcomes is a task goal and not an ego goal (de Charms, 1983). With task goals, students believe they are responsible for the outcome and they are informed of the mastery criteria for success. The teacher provides criterion-referenced performance feedback that emphasises mastery and progress rather than social comparison. Under such conditions, failure on a diagnostic test is more likely to be constructive rather than destructive. Diagnostic assessment therefore helps to foster the above conditions so that the perceived ability stratification among students in the class would be reduced. This will consequently help to improve the self-esteem and intrinsic motivation of low achievers in the group.

Diagnostic assessment, unlike norm-referenced assessment, does not encourage competitive structures in the classroom. In competitive structures, the success or failure of students is largely determined by their performance relative to other students i.e., social comparison is central to such structures. This tends to discourage low achieving students who failed to compete with their peers. It also discourages students from helping each other with their academic work, threatens peer relationships and encourage the segregation of high and low achieving students (Deutsch, 1979). Therefore competitive learning structures in the classroom discourages intrinsic motivation as well as encourages students to attribute success and failure to ability rather than to effort. This will be very harmful to the low achieving students.

On the other hand, there is a motivational aspect to the individualistic structures in a classroom applying the model of diagnostic assessment. Success is
based on criterion-referenced assessment. Although all students being assessed on the 
same objectives and using the same performance standards may give rise to a 
competitive structure, nevertheless competitiveness among the students are reduced. 
Moreover in diagnostic assessment emphasis is placed on monitoring each student's 
progress in mastery of intended objectives. Students are therefore more inclined to 
help each other, and success and failure on a test are more likely to be attributed to 
effort rather than ability. This generates conditions that support intrinsic motivation.

**Test anxiety.** The findings of this study suggested that diagnostic assessment 
is perceived as being anxiety provoking by the group. This is hardly surprising since 
test anxiety and fear of failure are aroused in an achievement task (Atkinson & 
Feather, 1966). Comparatively, the low achievers were more tense, felt more 
threatened and more fearful relative to the high achievers. This agrees with past 
 studies which have repeatedly shown substantial negative correlations between 
measures of test anxiety and performance on tests (Hill, 1984). The higher test anxiety 
of low achievers may be due to the expectation that the test may be difficult and the 
students' perceptions of good performance on the test to be particularly important 
since the test contributes to the overall grade of the student on the course. Causal 
attributions for failure and success based on past performance were hypothesised to 
relate to feelings of pride and shame and to both actual performance and expectations 
of future performance (Weiner, Frieze, Kukla, Reed, Rest, & Rosenbaum, 1971). The 
low achievers may be more test-anxious because they attribute failure on past tests to 
a lack of ability and to attribute success to effort. Nevertheless the positive 
perceptions of the low achievers toward diagnostic assessment suggested that the 
negative effects of text anxiety could have been worse. Past studies done by Hill &
Wigfield (1984) suggested that the debilitating effects of test anxiety may be reduced by informing students on the nature, difficulty and format of the test; setting test items that allow each student a reasonable level of success; and reducing emphasis on social comparison. These requirements are satisfied by diagnostic tests since they are based on known domains of instructional objectives while the interpretation of these tests are criterion-referenced and not norm-referenced.

**Student self-efficacy.** Diagnostic assessment offers an opportunity for low achievement students to have feelings of competency and self-efficacy. Bandura (1977) defines self-efficacy as a person's specific beliefs about his or her ability to perform certain actions or bring about intended outcomes. Self-efficacy judgments refer to expectations about personal competence in relation to a behavioural domain. Bandura (1982) have demonstrated that a person's self-efficacy determines how much effort he is willing to expend and how long he will persist in the face of obstacles and aversive experiences. He provided four sources of self-efficacy. Of interest here is the enactive information source. This information refers to direct experiences and accomplishments. The diagnostic assessment model used in this study sets a mastery criterion of 80% for the tests. The average difficulty of the tests is .54. Therefore these tests are achievement tasks that are optimal for the individual student i.e. of moderate difficulty and achievement-oriented. From a motivational point of view, it is very important that students experience regular positive feedback. Therefore it is recommended that when evaluating the learning outcomes of low achievers, individual rather than group criteria be used. Diagnostic assessment compares the outcomes of the test with the criterion of mastery of instructional objectives and not with the present outcomes of other students in the same group. This practice is crucial
for the less able students who might otherwise receive little positive feedback. There is strong evidence that self-efficacy is best enhanced if longer term goals are supported by a carefully sequenced series of sub-goals with clear criteria that students find attainable (Bandura & Schunk, 1981). These requirements are met by mastery learning procedures (Driscoll, 1986). In the context of diagnostic assessment, mastery of the diagnostic tests are the sub-goals while the final summative examination is the long term goal. Repeated success on the diagnostic test will help low achievers to develop less fear of failure and help improve their self-esteem and self-efficacy.

**Attributions for success and failure.** Past research by Weiner (1986) and Nicholls (1984), among others, have demonstrated that students' attributions for success or failure in learning tasks have a very significant influence on their motivation and behaviour. In his theory of achievement motivation, Weiner identified emotional consequences when success or failure is attributed to ability, effort, luck or task-difficulty. Success that is attributed to ability or effort leads to pride and self-esteem, failure attributed to lack of effort leads to guilt, and failure attributed to stable factors such as lack of ability or high task difficulty leads to hopelessness. According to Nicholls, many older children and adults conceive of ability as a stable trait that is judged normatively i.e., by comparing performances of different individuals. This conception promotes normative superior performance as an indicator of ability. Low achievers, based on past performance, typically perceive themselves as having insufficient ability to do well on subsequent summative classroom tests tend to display helplessness. They may not expend as much effort to learn as hoped by the teacher because they expect their efforts to result in failure thus reinforcing their low ability. On the other hand, high achievers tend to prefer tests of medium difficulty
which are likely to confirm their ability by distinguishing them from low achieving students who try to avoid such tests. Diagnostic assessment does not judge the ability of students by social comparison. It promotes the concept of ability as mastery of learning objectives through effort so that gains in content domain mastery are indicative of enhanced ability. Failure in a diagnostic test is then attributed to a lack of effort and not a lack of ability. The sense of shame when failing the test is thus reduced.

**Self-assessment.** Diagnostic assessment provides an external standard for self-assessment. This external standard takes the form of a criterion for mastery of instructional objectives of a given unit. The model provides the students with clear intended learning outcomes which they can use before and during instruction. It also provide for self assessment during and after completing the unit. Harter (1986) pointed out that positive learning experience, especially rewarding effortful accomplishments, help a student to establish adequate internal standards and to develop a stable identity as a successful learner. It is hoped that through self-assessment, students will gain self-regulatory control over the objective-directed learning in diagnostic assessment and become relatively independent of the teacher.

**Perceptions of Multiple-choice and Essay Formats**

College teachers typically use both multiple-choice tests and essays to assess course content. The choice of test format to be used for diagnostic tests is important to the test constructor because if students actually perceive a test format more positively than the other, there is a likelihood of enhancing optimal test motivation and minimising debilitating emotional reactions such as worry and test anxiety. Past studies have shown that test anxiety is related to test performance. For example,
Arkin, Kolditz and Kolditz, 1983 noted that task-irrelevant thoughts such as inadequacy, helplessness and concern over loss of status preoccupy the test-anxious student to the extent that performance is diminished. According to Williams (1991), it is this cognitive-interference or worry component of test anxiety that has been found most often to relate to test performance. The present study represents an attempt to assess the perceptions, emotional reactions and affective dispositions with respect to various critical facets of both formats.

The results show a significant difference between the perceptions of students towards multiple-choice and essay format tests. The strong preference of students for multiple-choice over essay type formats agrees with the findings of Zeidner (1988).

**Test anxiety.** The results with regard to the relationships between test formats and test anxiety were consistent with those obtained by Zeidner. In the present study, the facets ‘fear of failure’ and ‘anxiety’ were used to assess the test anxiety of students on the two formats while only the facet ‘anxiety’ was used in Zeidner’s research. The findings showed that the multiple-choice format test evokes significantly less anxiety and less fear of failure for the group, low achievers and high achievers. However only the ‘fear of failure’ of high achievers was significantly lower on the multiple-choice format as compared to the essay format.

The higher test anxiety on multiple-choice format tests may be explained by the attributional theory proposed by Weiner et al. (1971). Attribution theory may be relevant to an understanding of the nature of the high test anxiety on the essay format test. The students may have focus on reasons for their past performance on essay tests e.g., inadequate writing skills and lack of ability in expressing themselves. These task-
irrelevant thoughts may have preoccupy the test-anxious low achieving student to the extent that performance is diminished.

**Interpretation of Observed Test Scores for Diagnostic Decisions**

The use of item-analysis is an important component in the diagnostic assessment model implemented in this study. Items that discriminates well between masters and non-masters in a test provide information for the teacher to make diagnostic decisions on the effectiveness of teaching as well as student learning of objectives. Hence an item that discriminates perfectly between the masters and non-masters in a test can focus the teacher's attention to the objective of the item so that remedial instruction or corrective feedback can be given.

With a perfectly discriminating item, all students should be classified in cells 1 and 4 i.e., all test masters should answer the item correctly, and all non-masters fail to answer it correctly (Table 21).

<table>
<thead>
<tr>
<th>Test classification</th>
<th>Master</th>
<th>Non-Master</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Item classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-master</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

**Table 21**

Matrix to Classify Item and Test Mastery Distinctions
Such perfectly discriminating items were absent in the four multiple-choice tests given in this study. The researcher therefore needs to develop ground rules for making decisions about what was an acceptable deviation from it. In this study two approaches were used. One was to use statistics which would described the observed distribution. The other was to inspect the raw observed test data to pick out items which seem to deviate from the rest of the set.

**Interpretation using Brennan discrimination index, B.** For the first approach, three discrimination indices $B$, $r_{pb}$ and $\phi$ were used on the observed scores for any given item. Of interest is whether these three indices were able to highlight items which deviated from the function of discriminating between masters and non-masters. The higher the values of these indices, the more the item fits the description of a perfectly discriminating item. The Brennan discrimination index $B$ for an item indicates the difference between the proportion of students getting the item correct in the mastery and non-mastery group. Due to its ease of interpretability, it is the discrimination index of choice for diagnostic decisions using the multiple-choice diagnostic tests given during treatment. The diagnostic tests are designed to provide specific information on individual learning deficiencies and misunderstandings. When these tests are interpreted carefully they can provide information to students and teachers to regulate learning processes. In this study the rule used is that items with $B > .30$ and $p > .60$ are considered to be items that discriminates satisfactorily between masters and non-masters in the test. Diagnostic information is also gathered by inspection of the raw observed test data.
An item with $B = 0$ indicates a non-discriminating item with a difficulty index equal to 1. Although such an item tells us nothing about individual differences among students, this item may reflect an important objective and is acceptable in a criterion-referenced test.

Among the four multiple-choice tests, there were two items in Test 4 shown in Table 16(a) and one item in Test 5 shown in Table 16(b) with $B = 0$. The proportions of students in the mastery and non-mastery groups that obtain these items correct are equal to 1. This means that all the students get the item correct (difficulty index = 1) and the item seems to be equally effective for both groups. However, if both proportions are low, this will mean that most of the students are not answering the item correctly (low difficulty index). If this were the case, the teacher will need to find the probable cause of the failure to acquire the objective assessed by the item in order to correct incomplete prior learning. Alternatively, the low difficulty may suggest a need for the non-discriminating item to be revised. For norm-referenced tests, this argument is not valid because such tests are designed to identify student differences.

There were two negatively discriminating items identified among the four multiple-choice tests. Item 5 ($B = -.04, p = .85$) in Test 4 shown in Table 16(a) and item 13 ($B = -.01, p = .06$) in Test 5 shown in Table 16(b). Popham and Husek (1969) argued that negatively discriminating items must be poor since the best students seem to get them wrong, while the poor students get them right; i.e., such items misidentify student differences. A more careful analysis shows that this argument does not apply here. Item 5 is an easy item that is almost non-discriminating with $B$ approximately equal to zero. Therefore, the item appears to be equally effective for both groups. Compared to item 5, item 13 is a difficult item that is almost non-discriminating. The majority of the students failed to answer this item correctly.
suggesting that either the item or the instruction of the objective of the item (or both) may need to be revised.

Except for the three non-discriminating and two negatively discriminating items identified in the tests, the rest are positively discriminating items. There are 21 out of 24 (87.5 %) such items in Test 4, 18 out of 20 (90 %) in Test 5, 16 out of 16 in Test 6 (100 %) and 23 out of 23 in Test 7 (100 %). According to Popham and Husek (1969) a positively discriminating item is acceptable since it identifies student differences in the desired direction. However, Brennan (1972) pointed out that a positively discriminating item results in one group of students (masters) performing less effectively than another group of students (non-masters). He argued that if instruction preceding the criterion-referenced tests that is equally effective for all students, then the distribution of incorrect responses on an item should be randomly distributed among the two groups. In this case, $B$ should be zero and not highly positive. Thus, a positive $B$ value may indicate that the instruction needs to be revised in order to be more effective for the non-mastery group. On the other hand, if equal effectiveness of instruction for both groups results in a positive $B$ value, the item may need revision.

In this study, items that have a $B > .30$ and $p > .60$ are taken to be ‘good’ items that discriminates effectively between the masters and non-masters. The non-mastery group of students face difficulties in achieving the objectives of these items. Corrective feedback is given to the whole class as a whole while an assignment on these objectives is given to the non-masters. In this study, remedial instruction was given for objectives represented by 10 out of 24 items in Test 4 (42 %), 12 out of 20 items in Test 5 (60 %), 4 out of 16 items in Test 6 (25 %) and 2 out of 23 items in Test 7 (9 %).
Interpretation by inspection of raw observed data. As a rough indicator of discriminating items, $B$ worked. However it appeared to lose valuable information which could be gained from inspection of the observed data. The $B$ value for item 6 (Test 4) is 0.56. This indicates that the percentage of masters who mastered the item exceed the percentage of non-masters who mastered the item by 56%. However it appears to lose valuable information regarding the percentage of masters who failed to master the item. An inspection of the raw data in Table 17 shows that for this item 33.33% of the group, are test masters who are classified as non-item masters i.e. about 75% of the test masters failed this item. This is the most inappropriate cell to be if an item is discriminating adequately, so $B$ does not seem to be working that well because it does not highlight the underlying distribution of masters in cells 1 and 3 in Table 21 in the same way as the raw observed data. Item 22 (Test 4) is again suspect for the same reason since 30.31% (in cell 3) of the group, almost three-quarters of the test masters are misclassified by this item. The $B$ value for this item is .35. The observed data therefore highlight the problem of misclassification of masters by item which cannot be indicated by the value of $B$. A high $B$ value could be due to the reasonable non-master distinction in cells 2 and 4, for example 100% and about 75% of the non-test masters failed item 6 and item 22 respectively. These two items have a large proportion of the group in cell 3 and therefore poses a problem for diagnosis. To fall into that cell a student must appear to be a master of the unit but fail to answer the item correctly. This is a difficult classification to obtain because the criterion for mastery of 80% or more should ensure that masters do not fail to answer many items. Other problematic items identified by observed data are item 9 (Test 5) with 24.24% in cell 3, item 13 (Test 5) with 51.52% in
cell 3. These items may need revision or point to ineffective instruction on the objective represented by these items.

The relative proportion of students in cells 2 and 4 was also looked into. A larger proportion of students in cell 2 than in cell 4 indicates that a higher percentage of the students passed the item correctly but were classified as non-masters compared to a lower percentage who failed the item and were classified as masters. For each of the items 4, 5, 8, 10 and 12 in Test 4, 21.21% of item-masters were classified as test non-masters while 3.03% failed the item and were classified as test non-masters. In Test 5, 39.39% of item-masters were classified as test non-masters while 6.06% failed the item and were classified as test non-masters. This may be of concern for diagnosis purpose because these items picked out only about one-eighth of the non-masters in the group. A possible explanation is that many students have some idea about the unit but are not sufficiently certain of the concepts involved to answer all the items correctly and hence did not attain the required mastery score.

When the two approaches for interpreting the observed data are compared, it is clear that B provides an adequate indicator of item discrimination. However valuable data may be lost from the interpretation if an inspection of the raw data is not performed.

Agreement Among Discrimination Indices

The three item discrimination indices $B$, $r_{pb}$ and $\phi$ were compared for the four multiple-choice tests in two different ways. The Pearson-product moment correlation $r$ was used to assess the agreement between pairs of indices for each of the tests. The findings indicated that there is a strong positive correlation between $B$ and $\phi$ ranging from .87 to .99 for the tests. The mean correlation of .93 suggests that both indices varies consistently irrespective of the difficulty levels of the tests. The phi coefficients used to assess whether
there is a strong correlation between pairs of indices used in selecting 'good' and 'poor' items for each test indicated that $B$ and $\phi$ has the closest agreement with a mean phi coefficient value of .78. This suggests that there is a strong relationship between the discrimination indices $B$ and $\phi$. Since the Brennan index $B$ is much easier to calculate, easy to interpret and very highly related to the phi coefficient estimate, it is preferred for teacher made mastery tests for use in item analysis to discriminate between 'masters' and 'non-masters'. Moreover, it is much easier to and quicker to determine the number of discriminations made by all items on a test using $B$ than it is to calculate the $\phi$ correlation coefficient for every item. The relationship between $r_{pb}$ and $B$ appears to be the weakest with a moderate mean correlation of .58 and a mean phi coefficient of .60. This is supported by Brennan (1970) who compared $B$ with $r_{pb}$ for six sets of synthetic data that generated different forms of distribution of test scores ranging from a normal distribution to a very negatively skewed distribution. In his analysis, a comparison of the values of $B$ and $r_{pb}$ for the test items showed that $r_{pb}$ consistently declared fewer items to be significant as the distribution of scores became more negatively skewed. The criterion-referenced tests used in this study generally produce distributions of scores that are negatively skewed. The strong correlations of .82 (Test 4) and .87 (Test 5) with normal distribution of test scores, weak correlations of .39 (Test 6) and .22 (Test 7) with skewed distribution of scores appears to agree with the findings of Brennan. The mean correlation of .58 between $B$ and $r_{pb}$ for the four tests shows that the discriminating power of an item differs for the two discrimination indices. Brennan's study also found that the number of significant items resulting from the use of $B$ was very much dependent upon the cut-off point or criterion-score chosen for differentiating between the mastery and non-mastery group. The present study did not look into this problem. The results
of this study suggest that \( B \) and \( r_{pb} \) should not be used interchangeably when the distribution of test scores is not normal. However there is a strong indication that \( B \) and \( \phi \) may be used interchangeably irrespective of the form of distribution of the test scores.

**Conclusions and Recommendations**

Many educators and test theorists believe that the integration of assessment and instruction can enhance learning (e.g., Nitko, 1989; Dochy, 1993). The use of diagnostic assessment as a potent educational tool that enhance instruction and learning has not been fully utilised by teachers in the classroom. However, it need to be stressed that the right dosage for integration of assessment into the teaching-learning process has yet to be determined (Dochy et al., 1996). Many of the findings of this study draw support from past studies of learning, thus increasing the confidence which I have in them.

**Importance of diagnostic assessment.** Classroom assessment affects students in many different ways. It guides their judgments of what objectives are important to learn, affect their motivation and self-concept, consolidates learning and affect the development of effective learning strategies. Diagnostic assessment is used by instructors to modify subsequent learning activities for a unit, prescribe remediation of group or individual deficiencies and provide on-going feedback to the students for the purpose of directing advanced learning. It appears to be a very potent force influencing learning.

Diagnostic assessment deserves very careful planning and considerable amount of time and effort from teachers. Hence, its development may be hampered by a lack of commitment and support from teachers and educators. Classroom assessment currently appear to receive less thought than other aspects of education. Its power to affect students is not widely perceived or discussed. A more professional approach to diagnostic assessment
would demand regular and thoughtful analysis by teachers of their personal assessment
practices and greater attention to the establishment of more consistent assessment policies,
procedures and criteria within and among educational institutions.

**Diagnostic assessment to assist learning.** The grading function of assessment has
been overly emphasised while its role in assisting students to learn has often been neglected.
The integral role of assessment in teaching and learning needs to be grasped, and its grading
and placement functions placed in their proper perspectives. In the Malaysian primary
school, it is hard to see any justification before the sixth year for placing much emphasis on
using classroom assessment for normative grading of student achievement, given the
evidence reviewed in the literature that norm-referenced assessment, with the social
comparison and competitive learning structure that accompany it, produces undesirable
consequences on the motivation of low achievers. Similarly, it is not beneficial to interpret
class tests normatively in the secondary schools until the third year before the placement of
students into the arts or science streams in upper secondary classes. At the college level,
normative grading of student achievement should be performed on completion of a course.

The undesirable effects of norm-referenced interpretation of tests, as compared to
criterion-referenced diagnostic assessment, include reduction of intrinsic motivation,
damaging test anxiety, ability attributions for success and failure that undermine student
effort, lowered self-efficacy for learning in the low achieving students, reduced use and
effectiveness of feedback to improve teaching and learning, encourage competitive and
causes poorer social relationships among the students.

An overemphasis on the grading function of assessment has also led to overuse of
features normally associated with standardised testing, such as very formal testing
conditions, speeded tests with rigid time limits, a restricted range of item formats, and emphasis on the overall score rather than the diagnostic function of assessment which is to identify strengths and weaknesses. According to Wood (1986), this may be appropriate in psychological testing, but are rarely appropriate in education testing.

It is recommended that much of the assessment activity in education might more profitably be criterion-referenced and directed solely to diagnosis, monitoring student progress and providing useful feedback to students, whereas the less frequent assessments for summative purposes should focus on describing student attainments. The likely improvement in the achievement scores and competence of students associated with diagnostic assessment would justify the higher investment of time and effort from educators who implement diagnostic assessment in their classrooms.

**Effective feedback.** One of the main features of diagnostic assessment is feedback to students. For feedback to be successful, it needs to follow certain guidelines. First, it should focus students' attention on their progress in mastering intended learning outcomes. Such emphasis on personal progress will enhance self-efficacy, encourage attributions of success and failures to effort, and divert attention from social comparison. The essence of this approach is captured by Easley and Zwoyer (1975), quoted in Crookes (1988):

> If you can both listen to children and accept their answers not as things to just be judged right or wrong but as pieces of information which may reveal what the child is thinking you will have taken a giant step toward becoming a master teacher rather than merely a disseminator of information. (p.25)
Second, feedback on students' answers to test items should be provided soon after students have completed a unit of instruction and took a criterion-referenced test consisting of items which sampled a set of intended outcomes. The students should be given opportunities subsequently to correct their errors using the feedback. These may include errors caused by a lack of prior knowledge, failure to discriminate correctly between concepts, and applying the wrong rules or principles in problem-solving. One of the strengths of diagnostic assessment is that, after an item-analysis, remedial feedback can be tailored to the kinds of misunderstanding mentioned. Students are then given subsequent opportunities to correct their learning deficiencies. Third, feedback should be related to need. Detailed feedback that contains elaborate information on guiding strategies that process the domain-specific information is reliably more helpful. If the need arises, remedial instruction on the same unit should be provided to help the student work through misconceptions or other weaknesses in performance. Fourth, feedback should promote motivational beliefs that value and support learning (Covington, 1992). Teachers should provide feedback that influence attributions or self-efficacy. Students should be given feedback containing motivational content describing the roles of effort and strategy use in successful performance.

**Diagnostic assessment and task-focused goals.** Numerous studies have focused primarily on two types of goals. Task-focused goals have the following characteristics: success defined as improvement, progress and mastery, value placed on effort, basis for satisfaction comes from mastery of task, intrinsic reasons for effort, evaluation criteria is absolute, errors viewed as informational and competence viewed as development through
effort. For of ability-focused goals success is defined as high grades relative to others, value is placed on avoiding failure, satisfaction comes from being the best, reason for effort is to demonstrate self-worth, evaluation criteria are norms and social comparisons, errors are viewed as evidence of lack of ability and competence is viewed as inherited and permanent (Ames, 1992; Dweck & Leggert, 1988; Anderman & Maehr, 1994). All the above features of task-focused goals are associated with criterion-referenced assessment rather than norm-referenced assessment. According to Maehr and Anderman (1993) the psychological environment of the classroom have a strong influence on whether task-focused goals or ability-focused goals are adopted by students. If the activities in a class emphasise relative performance and grades, then students are likely to adopt ability-focused goals. In contrast, students are likely to adopt task-focused goals if teachers stress on task-mastery, effort and progress. The use of diagnostic assessment in the classroom employs instructional practices that encourage students to adopt task-focused goals e.g., mastery of learning outcomes, monitoring of progress and absolute performance standards. The teacher avoids instructional practices that emphasise getting the highest grades or highest relative achievement on tests, social comparison among students and viewing failure as evidence of lack of ability rather than effort. However some studies on school climate suggest that the school or college as a whole can influence the goals that students adopt (Maehr & Midgley, 1991). Schools and colleges that emphasise different goals will have an effect on individual students’ goals and motivation. A school that places too much emphasis on grades and performances will likely create an environment that encourages students compete for rewards and focus on
grades as the main focus of learning. Such school-wide practices that stresses on ability-focused goals will interfere with classroom instructional practices.

Hence, the efforts of implementing criterion-referenced assessment in the classroom can be more fruitful if the school provides the necessary support and adopt the philosophy that learning should be for its own sake and that the role of effort and progress need to be appreciated. Otherwise the teacher's efforts to use task-mastery goals in the classroom may be undermined if the school or college as a whole emphasises relative performance, competition and extrinsic rewards. It is recommended that an evaluative study be done to identify current classroom and school-wide practices, procedures and policies that stressed task- or ability-focused goals. Efforts can then be taken to change existing practices that promotes ability-focused goals and implement innovative practices designed to foster task-focused goals.

Diagnostic assessment and classroom learning approach. The effects of different classroom learning structures have been examined in many studies. In particular, the relative merits of competitive, individualistic and cooperative learning approaches have received considerable attention. The traditional classroom applies a competitive structure where norm-referenced assessment is used to compare the relative performance of students. Diagnostic assessment is based on an individualistic structure whereby students are rewarded on the basis of their own work independent of the work of other students. In such a structure there is no interdependence among students. Each student strives for mastery of a criterion alone. Although this learning approach reduces the negative interdependence among students as in competitive structures that accentuate
ability differences and threaten self-esteem of low achievers, nevertheless this group of students may still judge themselves as incompetent when they fail a mastery test. The rate of learning of low achievers are low. They need more help with their academic work. If a cooperative learning structure is implemented, students can work together particularly for more complex problem solving tasks so that the high achievers can compensate for the deficiencies of the low achievers. Such a learning structure will encourage active engagement of all students and stimulate helping behaviours within groups. Johnson, Maruyama, Johnson, Nelson and Skon (1981) reported that although competitive and individualistic structures seemed equally effective, cooperative structures generally produced higher achievement among students. This is supported by Ng (1998) who found that specifically designed task structures based on behavioural objectives can help to structure student interaction in cooperative group learning. Mastery of these objectives provides a measure of the learning outcomes. Ng’s study also reported a significant improvement in the sociology test scores as a result of cooperative group learning in her tutorials. It is therefore recommended that future studies look into the effectiveness of implementing diagnostic assessment in the classroom using a cooperative learning structure but with the inclusion of criterion-referenced assessment of the individual members.

**Setting standards.** Criterion-referenced tests used in diagnostic assessment emphasises on performance standards based on the mastery of a content domain i.e., simply what students ‘know’. These standards are set using ‘cut-off’ scores so that judgments about mastery or non-mastery of the content domain can be made based on the
observed test score distributions. However there is now a current trend towards the use of assessment for standards that emphasise on student performances i.e., what students can 'do'. Resnick and Tucker (1990) listed three types of performance that can be assessed: performance tasks, performance examinations and portfolios of students' work. Taylor (1994) suggested that each performance type can be used to reflect different aspects of standards e.g., performance examination to assess students' breadth of knowledge and portfolios to assess their depth of understanding in a subject area. Nevertheless in the context of diagnostic assessment, further studies will need to look into the question of how to use the information gathered to diagnose the learning weaknesses of students in the classroom.

**Implications of Diagnostic Assessment for Educational Reform**

In discussing the potential impact of replacing norm-referenced assessments with criterion-referenced assessments on education in Malaysia, the implications of the use of these assessments on a large scale in the school system need to be addressed. These two types of assessments are implemented differently. If policy makers are to make a choice regarding the assessment to put in place, those who will participate in the decision-making process will need to

(a) be knowledgeable about the assumptions underlying each assessment, (b) understand the purposes to be served by each assessment, and (c) consider the influences of each assessment on the structure of schools.

**Assumptions of the assessments.** The underlying assumptions of both types of assessment must be clearly understood by educators and policy-makers otherwise the
tensions between assessment for mastery of criteria and assessment for comparison of relative performance can result in failure to achieve the changes expected if criterion-referenced assessment is implemented on a larger scale. Once educators and policymakers understand the assumptions and the tensions, a choice must be made. Should instruments be designed to grade and compare students, or should an assessment system that indicates whether students are achieving intended outcomes be created? Should policymakers continue with the former assessment knowing that the current methods of standardised assessment can lead to deleterious consequences?

**Purposes of assessments.** Among policymakers and educators there are disagreement on the purposes of schools. Some believe that schools are supposed to discriminate between the brightest and the weakest students while others believe that the capacity and creativity of every student should be fostered. A choice between both types of assessment will require much public debate on the above issue. It is stated in the Seventh Malaysian Plan (1996-2000) that the focus of educational efforts is to enhance students’ academic performance, strive for excellence and increase the success rate of students. In other words, the focus of education in Malaysia is to raise standards for all students and to ensure that all students succeed. To achieve these aims it is stated that extensive monitoring and evaluation of students need to be performed while new teaching and learning approaches need to be developed. Diagnostic assessment using criterion-referenced tests appears to be the assessment model that will truly support these educational reform efforts.
The entrenched influences of norm-referenced assessment on our ways of thinking about learners and about our role as educators cannot be ignored. Ultimately, if criterion-referenced assessment is to be successful our beliefs about learners need to be changed. We must begin to agree with Bloom (1976) who claimed that individual differences in learning ability are unreal; that most students are quite capable of achieving learning equality and that the individual difference we see in student attainments are the result of conditions unrelated to students' capacity to learn. However it may be difficult to change beliefs about the innate abilities of students because students, parents, educators and policy-makers have gone through an educational system where nationally standardised achievement tests such as UPSR, PMR, SPM and STPM are built to optimally differentiate between examinees to compare them. In such examinations, items are developed and selected based on their value in discriminating between examinees with higher and lower test scores. These tests have affected our thinking about students' capacities to learn. Therefore norm-referenced tests designed to rank students will need to be phased out while a concerted application of the criterion-referenced assessment model will be necessary to make a significant impact on our long-held beliefs.

**Influence on school structure.** The implementation of a diagnostic assessment model using criterion-referenced tests will require significant and long-term changes in schools. The schools will be given the role of ensuring that students attain the criterion or performance standards. The question arises whether the existing educational systems can break free from the underlying assumptions about individual ability differences in
students and to make the changes needed. To be able to do this, they will need continual and sustained support from all parties that have a stake in the education of our children. Although this is not a simple task, if the assumptions of criterion-referenced model are accepted, it is likely that an educational system that reinforces this model can be created and maintained. These assumptions include design of a curriculum that focus on fewer but more central concepts, setting standards that most students can achieve, training educators to internalise the standards so that they can consistently judge student performances and diagnose their weaknesses in learning. If clear intended outcomes and performance standards for students at different levels of the school system can be set so that the schools can provide learning environments that help students meet those standards, then the likelihood of success will be enhanced. School administrations will have to create institutional structures that support teachers, and teachers will have to create learning structures and practices that foster learning for each individual learner. Policy-makers and other stakeholders will have to provide the necessary support to schools who implement criterion-referenced assessment while challenging schools that resist change.

Implications for teachers. The diagnostic assessment model suggests that essential instructional objectives are central to a discipline or subject. If teachers are to judge the effectiveness of their teaching or diagnose the learning weaknesses of students they will have to possess (a) subject-matter knowledge, (b) an understanding of the cognitive skills and processes that are central to the subject, (c) pedagogical and classroom management strategies that help students remedy their weaknesses, and (d)
knowledge of item-analysis techniques that include both logical and empirical reviews of the test data. Teachers will need to have good pedagogical content knowledge as well as thorough understanding of criterion-referenced test theories since they are expected to interpret test assessment data for making diagnostic, monitoring and attainment decisions in the classroom. Teachers will also have to be competent constructors and assessors of diagnostic instruments. Past studies suggest that most teachers have received almost no training in classroom assessment during professional preparation (Schafer & Lissitz, 1987). Therefore, in addition to ongoing professional development in their subject areas, teachers will need professional training in appropriate classroom assessment practices for diagnosis purposes.

In summary, if diagnostic assessment is to play its role in the bigger model of criterion-referenced assessment in the education system, difficult changes will need to be made. The schools will have to ensure that students achieve the performance standards expected. If we no longer accept differential student performance as an inevitable outcome of differences in traits or individual differences, we will need to focus our attention to those situations that prevent students from achieving standards, whether these be social and economic status or poor classroom instructional practices and management. Professional educators will have to define the important learning outcomes that are central to different levels of schooling. When implementing the diagnostic assessment model, we have to avoid the twin dangers of demanding instant results and the temptation of returning to normative comparison of students when the model is not immediately successful.
Finally it is hoped that more widespread implementation of diagnostic assessment can support all schools in reaching the goals espoused in the Seventh Malaysian Plan. Despite the inevitable resistance to change and the public debate that will occur, if diagnostic assessment can be implemented, it is the model that will most likely support real changes in our schools.