CHAPTER IV

RESULTS

Performance Data

Gain Score Analysis

The current study was conceived after the classes had been completed and all the test data except scores and standard deviations had been discarded so it was not possible to calculate test reliabilities for these tests. However, other reliability measures of the seven diagnostic tests administered during the treatment of diagnostic assessment as well as the first semester examination (Pretest) and second semester examination (Posttest) were evaluated using Cronbach's alpha to check for internal consistency. The mean scores, standard deviations , standard errors of the means (SEM) and the Cronbach's alpha for the pretest and posttest is as shown in Table 1.

Table 1

Means, Standard Deviations, SEM and Reliabilities for the Pretest and Posttest

	Test Statistics			
	M	SD	SEM	Cronbach's alpha
Pretest	60.48	26.51	4.62	.90
Posttest	64.94	20.83	3.63	.85

M = Mean, SD = Standard Deviation, SEM = Standard Error of the Mean.

The pretest and posttest scores for the three achievement groups is as shown in Figure 5. The mean posttest score (64.94) is about four points higher than the mean pretest score (60.48).

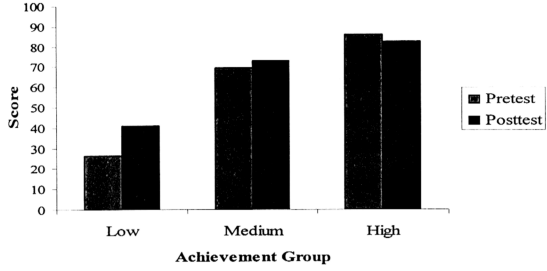


Figure 5. Pretest and posttest scores for three achievement groups.

The results of a gain-score analysis using a one tailed, correlated *t*- test for each of the ability groups, low (n = 11), medium (n = 11) and high (n = 11) achievers is as shown in Table 2.

Table 2

Gain Score Analysis for Each Achievement Group

	Pretest		Posttest		t-value	df
Ability	M	SD	M	SD		
Low	26.18	10.94	41.00	14.51	4.03**	10
Medium	69.18	6.51	72.73	12.46	0.84	10
High	85.91	4.04	82.45	5.43	-2.35*	10

M = Mean, SD = Standard Deviation.
* $p < .05$, one-tailed.
** $p < .01$

The results showed a significant gain in the mean scores of the low achievement group, $t(10) = 4.03$, $p < .01$; no significant gain on the medium group,

t (10) = 0.84, $p > .05$ and a significant difference in scores for the high group, t (10) = -2.35, $p < .05$.

Analysis of Variance

The mean posttest scores and standard deviations for the three achievement groups are reported in Table 3. The results of a one-way analysis of variance (ANOVA) on these data are shown in Table 4.

Table 3

Posttest Mean Scores and Standard Deviations for The Three Achievement Groups

Group	M	SD
Low	41.00	14.51
Medium	72.73	12.46
High	82.45	5.43

M = Mean, SD = Standard Deviation.

Table 4

One-way Analysis of Variance (ANOVA) For Posttest Mean Scores

Source	df	SS	MS	F
Between group	2	10338.97	5168.48	39.25**
Within group	30	3950.91	131.70	
Total	32	14289.88		

** $p < .01$

There was a significant main effect for the groups, $F(2,30) = 39.25$, $p < .01$. A post-hoc Tukey's HSD test indicated that there is a significant difference between the mean scores for low achievers and medium achievers as well as between low and high

achievers at an α - level of .05. Therefore the high and medium achievers maintained the same relative advantage over the low achievers in their physics scores after the treatment of diagnostic assessment.

Repeated Measures Data

The pretest-posttest design provides only two data points for each achievement group, the pretest score and the posttest score. Confounding variables may have affected these results. To lessen this problem, the test scores observed before and after the intervention of diagnostic assessment are analysed. If a trend emerge where the test scores increase over time after the treatment has begun, this will provide supportive evidence that the intervention is effective. However, the pre-treatment period of test scores need to form a stable baseline to provide a standard by which the effectiveness of the experimental intervention can be evaluated. Table 5 shows the mean pre- and post-treatment percentage scores for the three groups.

Table 5

Mean Pre-treatment and Post-treatment Scores for the Achievement Groups

	Low	Medium	High
Pre-scores			
1	30	57	67
2	26	58	68
3	16	60	73
4	25	61	77
5	26	59	74
6	26	69	86
Post-scores			
1	55	90	89
2	61	89	90
3	62	85	89
4	70	92	93
5	52	80	88
6	38	60	75
7	53	64	69
8	41	72	82

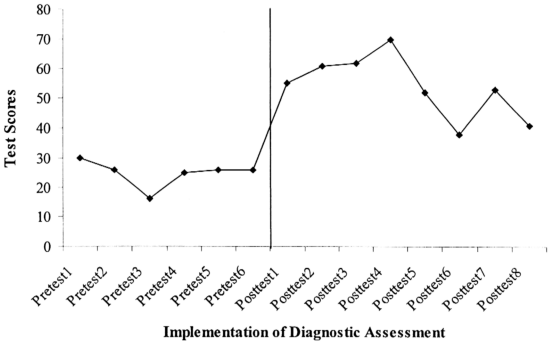
Figures 6(a), 6(b) and 6(c) shows pattern of test scores before and after the implementation of diagnostic assessment for the three achievement groups. The stable baseline and narrow range of variability for each group agree with the assignment of group members based on their first semester examination scores (Pre-score 6). The baseline period also function as a predictor for the level of the target achievement attained for each group after the intervention of diagnostic assessment.

An examination of the post-intervention baseline in Figure 6(a). shows a change at the point in which intervention is made – a change in level. There is also a change in slope or trend at the beginning of the intervention phase. This significant change in slope implies a change in the mean scores across the two phases. However there is no discernable upward or downward trend for the last four points. This baseline, however, is generally higher than the pre-treatment baseline.

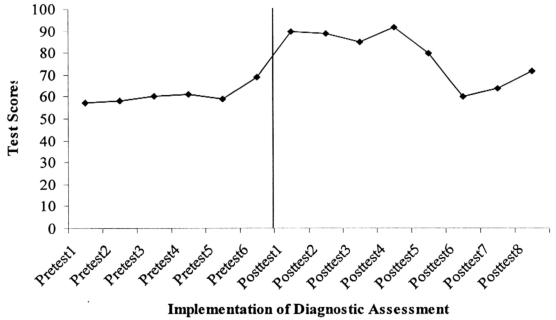
Figure 6(c) shows a post-treatment baseline that is a continuation of the steady trend established during the pre-treatment baseline measurement phase. There is a small change in level or slope across the two phases. The variability between the pre-treatment and post-treatment baselines is small. This suggested that the introduction of diagnostic assessment does not affect the high achievement group of students.

The post-treatment baseline in Figure 6(b). shows a change in level and slope at the beginning of the intervention phase. However the latter half of the baseline does not indicate a particular pattern.

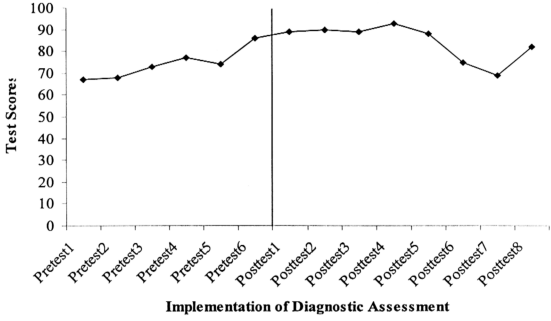
Figure 7 compares the pattern of test scores for the three achievement groups before and after the intervention of diagnostic assessment.



(a) Low Achievement Group



(b) Medium Achievement Group



(c) High Achievement Group

Figure 6. Pattern of test scores for three achievement groups using a time-series design

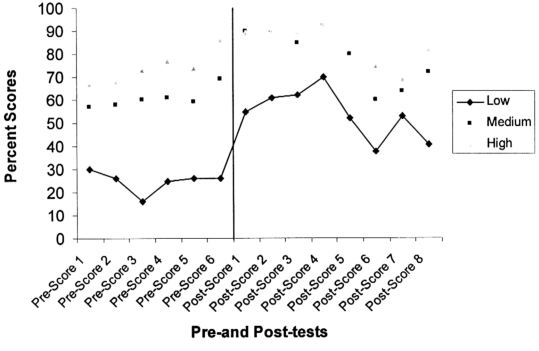


Figure 7. Comparison of the pattern of test scores for three achievement groups before and after intervention of diagnostic assessment

Judges' Ratings of Difficulty Level of Tests

Table 6 presents the mean and standard deviation of each rater's estimates of the difficulty levels of the first semester and second semester tests.

Table 6

Means and Standard Deviations for the Ratings

Rater	First Semester Tests		Second Semester Tests	
	Mean	S.D.	Mean	S.D.
1	2.75	0.83	2.29	0.70
2	2.25	0.43	2.86	0.64
Overall	2.50	0.61	2.43	0.68

Judges' ratings: not difficult, 1; somewhat difficult, 2; quite difficult, 3; very difficult, 4

The mean overall rating of difficulty is 2.50 for all the tests in the first semester and 2.43 for the tests in the second semester. These results suggest that there is no major differences in the level of difficulty of the tests for the two semesters. The judges concur with the instructor that the course content of Electricity and Magnetism in the first semester were more complex than Mechanics and General Physics due to the abstractness of the material. In the second semester, the modules on Waves, Quantum Physics, Atomic Physics and Nuclear Physics consists of abstract materials which may be a component of learning difficulty for students . Generally the difference in the complexity and difficulty of the course contents in the two semesters appears to be minimal.

Perceptions Data

The Cronbach's alpha reliability for the Likert scale used to assess the perceptions of students towards diagnostic assessment was found to have a value of .97. This reliability is considered good for group as well as individual comparison purposes. Higher scores (mean more than 3) suggest increasingly favourable perceptions whereas lower scores (mean below 3) suggest less favourable perceptions. Item ratings were linearly combined and the mean determined in order to form a composite perception score. This served as a global measure of perceptions. The mean ratings of the three achievement groups and the composite rating for the whole group on the scale is as shown by the chart in Figure 8.

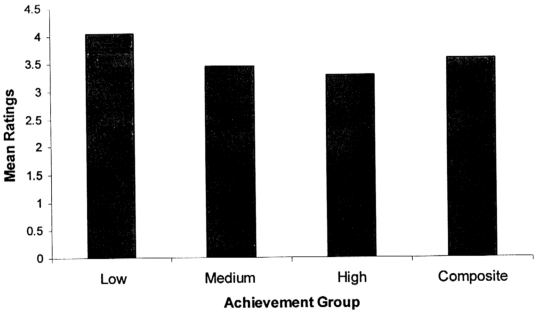


Figure 8. Bar-chart of mean and composite ratings on the Likert-scale for the three achievement groups

The results indicated that among the three groups, the low achievement group (mean, 4.05) has the most favourable perceptions of diagnostic assessment while the high group (mean, 3.29) has the least favourable perceptions. Overall the group

perceives the assessment favourably (mean, 3.6).

A one-way analysis of variance (ANOVA) conducted on these data was significant, $F(2,30) = 5.25$, $p < .05$. The results are presented in Table 7. The results of Tukey's HSD indicated that the only difference between the groups was between the high achievement group and the low achievement group.

Table 7

Analysis of Variance to Compare Mean Scores on Likert Scale for Three Groups

Source	df	SS	MS	F
Between group	2	3.53	1.76	5.25*
Within Group	30	10.08	0.34	
Total	32	13.60		

* $p < .05$

Perceptions Data Using Semantic Differential Scale

The groups were compared on both individual scales as well as on each of the five semantic differential dimension scores (i.e., mean of scale scores representing a given dimension). Mean scale ratings of more than 4 denote a positive perception towards diagnostic assessment while ratings less than 4 suggest a negative perception.

Table 8 shows the means and standard deviations for the semantic differential ratings of classroom diagnostic assessment. The data is presented in Figure 9. The mean and composite scores on the five dimensions is shown in Figure 10.

Table 8

Means and Standard Deviations for Semantic Differential Ratings on Classroom Diagnostic Assessment (n = 33)

Dimensions and Scales	M	SD
COMPOSITE	**4.16**	**0.14**
EVALUATION	**4.63**	**0.16**
Good/Bad	4.94	1.14
Valuable/Worthless	4.58	1.00
Pleasant/Unpleasant	4.58	1.06
Fair/Unfair	4.58	0.97
Informative/Non-informative	4.45	1.03
POTENCY	**4.05**	**0.16**
Strong/Weak	4.24	0.90
Light/Heavy	4.15	0.83
Soft/Hard	3.76	1.06
ACTIVITY	**4.16**	**0.42**
Fast/Slow	3.55	1.12
Hot/Cold	4.15	0.76
Active/Passive	4.73	0.84
Dynamic/Static	4.21	0.89
SUBJECTIVE ANXIETY	**3.53**	**0.20**
Relaxing/Tense	3.64	1.17
Evoke Fear/Not fear evoking	3.70	1.13
Threatening/Non-threatening	3.24	1.03
COMPREHENSIBILITY	**4.15**	**0.19**
Easy/Difficult	4.03	1.26
Simple/Complex	4.00	1.20
Clear/Confusing	4.42	0.84

M = Mean, SD = Standard Deviation.

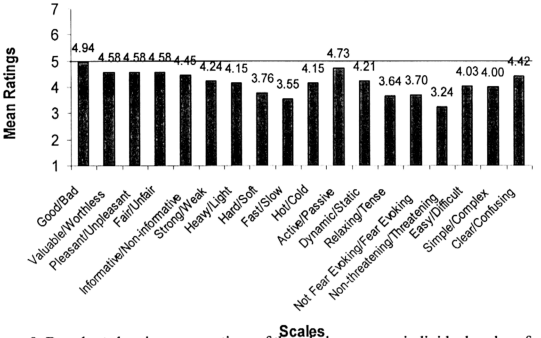


Figure 9. Bar-chart showing mean ratings of the whole group on individual scales of the semantic differential instrument

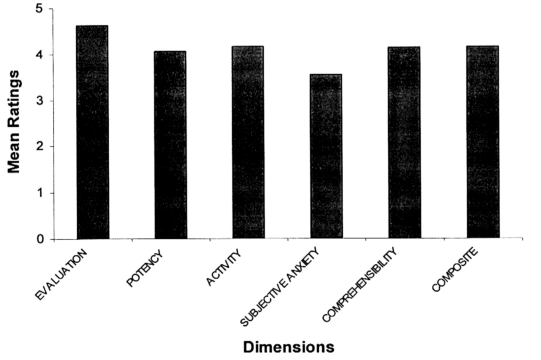


Figure 10. Bar-chart of composite and mean ratings of the whole group on the 5 dimensions of the semantic differential instrument.

Table 9 shows the means and standard deviations of the semantic differential ratings on each of the 18 scales and 5 dimensions for the low and high achievement groups. The data is presented as barcharts in Figure 11 and Figure 12. The Cronbach's alpha reliability estimates computed for the low and high achievement group responses to the semantic differential instrument are .75 and .73 respectively.

Table 9

Means and Standard Deviations of Semantic Differential Ratings for the Low and High Achievement Groups

	Low		High	
	M	SD	M	SD
COMPOSITE	**4.08**	**0.47**	**4.23**	**0.15**
EVALUATION	**4.76**	**0.32**	**4.44**	**0.32**
Good/Bad	5.27	1.27	4.64	0.81
Valuable	4.82	1.17	4.45	0.69
Pleasant/Unpleasant	4.64	1.12	4.64	0.92
Fair/Unfair	4.82	1.17	4.27	0.79
Informative/Non-informative	4.27	1.27	4.18	0.87
POTENCY	**4.18**	**0.22**	**4.00**	**0.26**
Strong/Weak	4.45	0.82	4.09	0.83
Heavy/Light	4.18	0.98	4.27	0.65
Hard/Soft	3.91	1.30	3.64	0.92
ACTIVITY	**4.14**	**0.60**	**4.30**	**0.36**
Fast/Slow	3.18	0.87	4.09	1.22
Hot/Cold	4.18	0.75	4.09	0.83
Active/Passive	4.82	0.87	4.91	0.54
Dynamic/Static	4.36	0.81	4.09	0.83
SUBJECTIVE ANXIETY	**3.30**	**0.31**	**4.12**	**0.22**
Relaxing/Tense	3.18	0.87	4.36	1.29
Not fear evoking/Fear evoking	3.73	1.01	4.18	1.17
Non threatening/Threatening	3.00	0.89	3.82	1.25
COMPREHENSIBILITY	**4.03**	**0.11**	**4.30**	**0.17**
Easy/Difficult	3.91	1.30	4.18	1.33
Simple/Complex	4.00	1.10	4.18	0.98
Clear/Confusing	4.18	1.08	4.55	1.37

M = Mean, SD = Standard Deviation

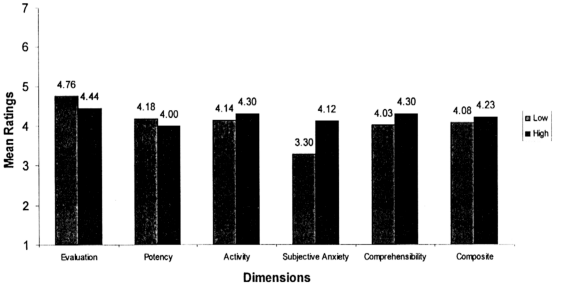


Figure 11. Barchart showing the mean ratings of the low and high achievement groups on the 5 scales in the 5 dimensions of the semantic differential instrument

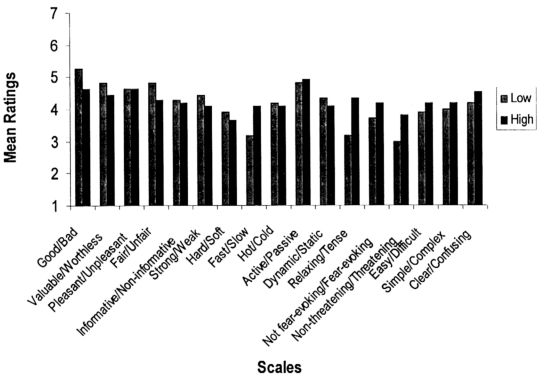


Figure 12. Barchart showing the mean ratings of the low and high achievement groups on individual scales of the semantic differential instrument.

The whole group was observed to perceive the classroom diagnostic assessment favourably (mean, 4.16) on all the dimensions. The Cronbach's alpha reliability estimate for the instrument is 0.75. The mean rating for the evaluation dimension alone was found to be higher at 4.63. An examination of intercorrelations between items in the evaluation dimension using the Pearson product-moment correlation correlations between .55 to .78. This moderate to high correlations lend support to the items included in this dimension. The Cronbach's alpha reliability estimate for the evaluation dimension alone is .90. Table 10 shows the correlation matrix for items in the evaluation dimension of the semantic differential scale.

Table 10

Correlation Matrix for Items in the Evaluation Dimension of the Semantic Differential Scale

	Good	Valuable	Pleasant	Fair	Informative
Good	1.00	0.74	0.60	0.68	0.63
Valuable	0.74	1.00	0.59	0.78	0.71
Pleasant	0.60	0.59	1.00	0.61	0.55
Fair	0.68	0.78	0.61	1.00	0.67
Informative	0.63	0.71	0.55	0.67	1.00

The favourable ratings of the group of students toward diagnostic assessment as measured using the Likert scale and the evaluation dimension of the semantic differential scale provide concurrent validity on the positive perceptions of students towards diagnostic assessment.

Additional comparisons of the ratings of the low and high achievement groups by individual scales on the evaluation dimension reveal consistently higher mean ratings for the low achievement group for four out of the five scales appearing on the

inventory. Specifically, the low achievement group, relative to the high achievement group, perceives diagnostic assessment to be 'good' (5.27 > 4.64), 'valuable' (4.82 > 4.45), 'fair' (4.82 > 4.27) and 'informative' (4.27 > 4.18). This comparison between the ratings on the evaluation dimension can be seen in the barchart as shown in Figure 12. Overall, the mean composite score on the evaluation dimension is higher for the low achievement group (4.76) compared to the high achievement group (4.44). This agrees with the results of the Likert-scale which indicates that the low achievement group perceive diagnostic assessment more favourably than the high achievement group.

At the same time however, diagnostic assessment is perceived as being more anxiety provoking by the low achievement group relative to their high achievement counterparts as shown by their ratings on the subjective anxiety dimension (3.30 < 4.12). On this dimension, the low achievement group, relative to the high achievement group, scored consistently lower in the 3 scales of 'tense' (3.18 < 4.36), 'threatening' (3.00 < 3.82) and 'fear evoking' (3.73 < 4.18). However overall, both groups reported unfavourable ratings for this dimension. This indicates that the classroom test situation is an anxiety evoking experience for both high and low achieving students.

Data on Essay and Multiple-choice Format Tests

The tests administered during the treatment of diagnostic assessment consisted on three essay-type tests and four multiple-choice format tests. Students were instructed to rate the features of the test formats in the Likert-scale based on their experiences on the seven tests and not on prior classroom testing experiences. The Cronbach's alpha reliability estimate, calculated separately for ratings of multiple-choice and essay diagnostic tests were .57 and .62 respectively, and was thus

considered to be satisfactory for group comparison purposes. Ratings of more than 3 suggest increasingly favourable test perceptions whereas scores below 3 indicate less favourable perceptions.

Table 11. shows the means and standard deviations for the composite score and each of the individual scales assessing both essay and multiple-choice format items. The data is presented in Figure 13.

A one-tailed correlated group t – test, comparing the mean composite perception ratings for essay ($M = 2.80$) versus multiple-choice ($M = 3.02$) test formats, was shown to be significant, $t\,(32) = 2.41$, $p < .05$, indicating that students perceive multiple-choice format tests more favourably than essay type tests.

Additional comparisons of the two formats by individual scales (Table 10), show significant difference on six of the fourteen facets of the classroom diagnostic tests appearing in the inventory. Specifically, the multiple-choice format, relative to the essay format, is perceived to be significantly 'easier' ($2.76 > 2.30$), 'more comfortable' ($3.24 > 2.39$), 'less fear of failure' ($2.70 > 2.18$), 'evoking less anxiety' ($2.55 > 2.18$), and 'generating higher expectancy of success' ($3.33 > 2.73$). However the essay-type test is perceived to be significantly 'more tricky'.

Table 11

Means and Standard Deviations for Likert-scale Ratings of Essay versus Multiple-choice Format Tests (n = 33)

	Multiple-choice		Essay		t-value
Items	M	SD	M	SD	
Composite Scale	3.02	0.39	2.80	0.44	2.41*
Difficulty	2.76	1.06	2.30	1.13	2.17*
Complex	2.73	1.07	2.42	0.97	1.47
Appropriateness	3.18	0.81	3.24	0.79	-0.37
Clarity	3.30	1.02	3.12	0.96	0.71
Valuable	3.21	0.65	3.42	0.75	-1.65
Tricky	2.09	0.98	2.55	1.09	-1.79*
Motivation	3.24	0.90	3.24	0.90	0.00
Comfortable	3.24	0.87	2.39	0.97	3.55**
Fairness	3.06	0.75	2.91	1.23	0.58
Fear of Failure	2.70	1.13	2.18	1.16	2.19*
Interesting	3.21	0.89	3.15	1.12	0.25
Anxiety	2.55	1.12	2.18	1.13	1.88*
Effectiveness	3.67	0.89	3.36	0.86	1.44
Success	3.33	0.85	2.73	0.94	2.79**

M = Mean, SD = Standard Deviation

* $p < .05$, one –tailed.
** $p < .01$

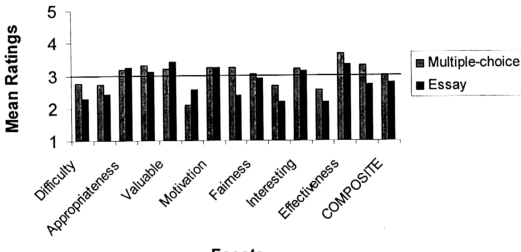


Figure 13. Means for Likert-scale ratings of essay versus multiple-choice format tests

The mean ratings and standard deviations for the 14 facets of the tests for the low and high achievement groups are shown in Tables 12(a) and 12(b).

Table 12

Mean Ratings and Standard Deviations for the Low and High Achievement Groups (n = 11) on the Likert-scale

Facets	Multiple-choice		Essay		*t*-value
	M	SD	M	SD	
Composite	2.98	0.47	2.78	0.42	
Difficulty	2.91	1.30	2.45	1.44	
Complex	3.00	1.10	2.36	1.03	3.13**
Appropriateness	2.73	0.65	3.27	0.47	-2.63*
Clarity	2.91	0.83	3.00	1.00	
Valuable	3.18	0.75	3.00	0.89	
Tricky	1.82	0.87	2.36	1.12	
Motivation	3.45	0.82	3.45	1.13	
Comfortable	3.18	0.98	2.45	1.04	2.67*
Fairness	3.27	0.91	3.09	0.94	
Fear of Failure	2.91	1.22	2.36	1.43	
Interesting	3.27	1.01	3.36	1.20	
Anxiety	2.18	0.87	2.09	0.94	
Effectiveness	3.64	1.03	2.91	0.83	2.03*
Success	3.27	1.01	2.73	1.35	

M=Mean, SD=Standard Deviation.

*p < .05, one-tailed.
** p < .01

(a) Low Achievement Group

Facets	Multiple-choice		Essay		*t*-value
	M	SD	M	SD	
Composite	3.14	0.42	2.69	0.61	
Difficulty	2.64	0.92	1.82	0.60	2.76*
Complex	2.64	0.92	2.00	0.63	2.28*
Appropriateness	3.45	0.93	3.18	1.17	
Clarity	3.64	1.12	3.09	1.14	
Valuable	3.27	0.47	3.64	0.51	-2.39*
Tricky	2.55	1.21	2.64	0.92	
Motivation	3.09	0.83	3.09	0.83	
Comfortable	3.55	0.69	2.09	0.83	4.28**
Fairness	2.64	0.51	2.73	1.49	
Fear of Failure	2.91	1.14	2.00	0.89	2.09*
Interesting	3.09	0.83	2.73	1.01	
Anxiety	3.09	1.22	2.36	1.21	
Effectiveness	3.91	0.54	3.91	0.54	
Success	3.55	0.69	2.36	0.51	4.49**

M=Mean, SD=Standard Deviation
* $p < .05$, one-tailed.
** $p < .01$

(b) High Achievement Group

Test Item Analysis Data

In this study, seven diagnostic tests were administered to the students during the implementation of diagnostic assessment in the classroom. Table 13 shows the test format, test length , test statistics and test reliabilities for these tests. Table 14 shows the Pearson product-moment correlation between test scores in Physics and the second semester examination scores for Further Mathematics and Malaysian Studies. The correlations between the Physics and Further Mathematics scores range between .48 to .87 with a mean of .71 while that between Physics and Malaysian Studies varies from .35 to .67 with a mean of .56. The strong correlations between Physics and Further Mathematics scores indicate agreement in measuring the same construct while the moderate correlations between Physics and Malaysian Studies scores indicate

suggest that these scores may not measure the same construct. These correlations provide supportive evidence of construct validity for the diagnostic tests.

Table 13

Test Format, Test Length, Test Statistics and Test Reliabilities

Test	Format	p	Length	Mean	SD	SEM	Cronbach's alpha	Inter-rater reliability	Test-Exam Correlation
1	Essay	.64	9	24.36	5.33	2.13	.84	.89	.64
2	Essay	.67	13	15.95	3.00	1.34	.80	.87	.82
3	Essay	.64	4	23.61	4.80	2.49	.73	.86	.67
4	MC	.76	24	20.27	3.20	1.53	.77	.88	.72
5	MC	.55	20	14.67	3.81	1.66	.81	.85	.81
6	MC	.30	16	9.18	3.46	1.76	.74	.94	.58
7	MC	.21	23	14.27	3.61	1.94	.71	.83	.41

p = Test difficulty index, MC = multiple-choice

Table 14

Correlations of Physics Test Scores with Further Mathematics and Malaysian Studies

Physics	Further Mathematics	Malaysian Studies
Test 1	.72	.59
Test 2	.78	.67
Test 3	.76	.60
Test 4	.69	.58
Test 5	.87	.64
Test 6	.70	.47
Test 7	.48	.35

The discrimination index used for the essay format test is the item-total correlation using the Pearson product-moment correlation coefficient, r. For the multiple-choice tests, three different discrimination indices are used. These are the Brennan index B, the point-biserial correlation coefficient r_{pb} and the phi coefficient

ϕ. Table 15 shows the item-total correlations for the essay format Tests 1, 2 and 3. The mean value of r for Test 1 is .67, for Test 2 is .52 and for Test 3 is .72.

Table 15

Item-total Correlations for Essay Format Tests

	Item-total Correlation, r		
Item	Test 1	Test 2	Test 3
1	.31	.45**	.36*
2	.63**	.43*	.88**
3	.65**	.00**	.74**
4	.81**	.58**	.88**
5	.62**	.79**	
6	.76**	.60**	
7	.74**	.77**	
8	.66**	.51**	
9	.81**	.66**	
10		.52**	
11		.51**	
12		.32	
13		.78**	

* $p < .05$, 2-tailed
** $p < .01$

A summary of the item statistics of difficulty index p, Brennan index B, point- biserial correlation r_{pb} and phi coefficient ϕ for Tests 4, 5, 6 and 7 is shown in Tables 16(a), 16(b), 16(c) and 16(d). Generally, the mean values of p, r_{pb}, B and ϕ for the tests satisfy the criterion of a 'good' item i.e. $.6 < p < 1.0$ and B, r_{pb}, $\phi > .3$.

Table 16(a)

Item Difficulty and Discrimination Indices
for Tests 4

	Item Statistics			
Item	*p*	*B*	r_{pb}	ϕ
1	.94	.25	.58**	.45**
2	.91	.21	.40	.31
3	1.0	.00	.00	.00
4	.97	.13	.36	.31
5	.85	-.04	.07	-.04
6	.42	.56	.67**	.49**
7	.94	.25	.58**	.45**
8	.94	.09	.22	.15
9	.73	.47	.59**	.45**
10	.97	.13	.12	.31
11	1.0	.00	.00	.00
12	.94	.09	.24	.15
13	.85	.30	.41**	.18
14	.91	.38	.53**	.56**
15	.73	.30	.36	.29
16	.88	.34	.30	.44**
17	.85	.46	.68**	.55**
18	.79	.55	.63**	.57**
19	.88	.46	.44**	.55**
20	.91	.38	.63**	.56**
21	.91	.21	.35	.31
22	.48	.35	.31	.30
23	.88	.30	.30	.35*
24	.61	.64	.50*	.56**
Mean	**.85**	**.29**	**.40**	**.34**

*p < .05, 2-tailed
**p < .01

Table 16(b)

Item Difficulty and Discrimination Indices
for Test 5

	Item Statistics			
Item	p	B	r_{pb}	ϕ
1	1.0	.00	.00	.00
2	.94	.08	.15	.13
3	.88	.27	.65**	.41**
4	.73	.11	.11	.12
5	.67	.61	.84**	.65**
6	.82	.34	.58**	.42**
7	.82	.40	.77**	.52**
8	.76	.17	.43	.19
9	.48	.02	-.22	.02
10	.70	.54	.65**	.59**
11	.73	.48	.43	.53**
12	.67	.73	.70**	.77**
13	.06	-.01	.15	-.02
14	.79	.47	.70**	.57**
15	.82	.40	.56*	.52**
16	.91	.08	.25	.13
17	.67	.49	.79**	.53**
18	.70	.54	.88**	.59**
19	.79	.34	.50*	.42*
20	.76	.34	.42	.42*
Mean	**.74**	**.32**	**.47**	**.38**

* $p < .05$, 2-tailed
** $p < .01$

Table 16(c)

Item Difficulty and Discrimination Indices
for Test 6

	Item Statistics			
Item	*p*	*B*	r_{pb}	ϕ
1	.64	.38	.26	.36*
2	.61	.57	.69**	.53**
3	.67	.48	.49	.47**
4	.58	.47	.46	.43
5	.49	.31	.53*	.28
6	.49	.45	.48	.42
7	.52	.41	.53*	.38
8	.46	.35	.54*	.33
9	.67	.33	.34	.33
10	.24	.08	.30	.09
11	.82	.26	.76**	.31
12	.49	.27	.34	.24
13	.91	.13	.45	.21
14	.55	.37	.46	.34
15	.52	.27	.28	.24
16	.58	.28	.23	.26
Mean	**.58**	**.34**	**.45**	**.33**

* $p < .05$, 2-tailed
** $p < .01$

Table 16(d)

Item Difficulty and Discrimination Indices
for Test 7

	Item Statistics			
Item	p	B	r_{pb}	ϕ
1	.36	.63	.46*	.53**
2	.70	.38	.38	.34*
3	.18	.49	.35	.52**
4	.36	.63	.46**	.53**
5	.79	.27	.46**	.27
6	.76	.13	.26	.12
7	.73	.35	.60**	.32
8	.91	.12	.49**	.16
9	.42	.37	.36	.30
10	.97	.04	.11	.09
11	.15	.35	.13	.40*
12	.58	.36	.34	.30
13	.79	.27	.49**	.27
14	.91	.12	.32	.16
15	.82	.23	.71**	.23
16	.82	.19	.28	.22
17	.58	.14	.29	.11
18	.36	.26	.36	.22
19	.55	.40	.26	.32
20	.48	.43	.41	.36
21	.91	.12	.20	.16
22	.88	.15	.44*	.19
23	.27	.52	.27	.46**
Mean	**.62**	**.30**	**.37**	**.29**

*p < .05
**p < .01

Item Data for Diagnostic Decisions

Tables 17(a), 17(b), 17(c), 17(d) show a summary of the item data for Tests 4, 5, 6 and 7 . An inspection of the percentage of the group who (a) pass test and item, (b) pass test and fail item, (c) fail test and pass item, and (d) fail test and fail item, as

shown in the relevant cells will provide valuable information on the discriminating property of the item. The value of B for each item is included to assess whether valuable information will be lost if B is used as the index of discrimination without inspection of the distribution of masters and non-masters into the four classes.

Table 17(a)

Item Data for Test 4

	Test	p	B	% pass test	% fail test		Test	p	B	% pass test	% fail test
Item						Item					
1	Pass	0.94	0.25	75.76	18.18	13	Pass	0.85	0.30	69.70	15.15
	Fail			0.00	6.06		Fail			6.06	9.09
2	Pass	0.91	0.21	72.73	18.18	14	Pass	0.91	0.38	75.76	15.15
	Fail			3.03	6.06		Fail			0.00	9.09
3	Pass	1.00	0.00	75.76	24.24	15	Pass	0.73	0.30	60.61	12.12
	Fail			0.00	0.00		Fail			15.15	12.12
4	Pass	0.97	0.13	75.76	21.21	16	Pass	0.88	0.34	72.73	15.15
	Fail			0.00	3.03		Fail			3.03	9.09
5	Pass	0.85	-0.04	63.64	21.21	17	Pass	0.85	0.46	72.73	12.12
	Fail			12.12	3.03		Fail			3.03	12.12
6	Pass	0.42	0.56	42.43	0.00	18	Pass	0.79	0.55	69.70	9.09
	Fail			33.33	24.24		Fail			6.06	15.15
7	Pass	0.94	0.25	75.76	18.18	19	Pass	0.88	0.46	72.73	12.12
	Fail			0.00	6.06		Fail			3.03	12.12
8	Pass	0.94	0.09	72.73	21.21	20	Pass	0.91	0.38	75.76	15.15
	Fail			3.03	3.03		Fail			0.00	9.09
9	Pass	0.73	0.47	63.64	9.09	21	Pass	0.91	0.21	72.73	18.18
	Fail			12.12	15.15		Fail			3.03	6.06
10	Pass	0.97	0.13	75.76	21.21	22	Pass	0.48	0.35	45.45	6.06
	Fail			0.00	3.04		Fail			30.31	18.18
11	Pass	1.00	0.00	75.76	24.24	23	Pass	0.88	0.3	69.70	15.15
	Fail			0.00	0.00		Fail			6.06	9.09
12	Pass	0.94	0.09	72.73	21.21	24	Pass	0.61	0.64	57.58	3.03
	Fail			3.03	3.03		Fail			18.18	21.21

Table 17(b)

Item Data for Test 5

	Test	p	B	% pass test	% fail test		Test	p	B	% pass test	% fail test
Item						Item					
1	Pass	1.00	0.00	54.55	45.45	13	Pass	0.06	-0.01	3.03	3.03
	Fail			0.00	0.00		Fail			51.52	42.42
2	Pass	0.94	0.08	51.52	39.39	14	Pass	0.79	0.47	54.55	24.24
	Fail			3.03	6.06		Fail			0.00	21.21
3	Pass	0.88	0.27	54.55	33.33	15	Pass	0.82	0.40	54.55	27.27
	Fail			0.00	12.12		Fail			0.00	18.18
4	Pass	0.73	0.11	42.42	30.30	16	Pass	0.91	0.08	51.52	39.39
	Fail			12.12	15.15		Fail			3.03	6.06
5	Pass	0.67	0.61	51.52	15.15	17	Pass	0.67	0.49	48.48	18.18
	Fail			3.03	30.30		Fail			6.07	27.27
6	Pass	0.82	0.34	51.52	27.27	18	Pass	0.70	0.54	51.52	18.18
	Fail			3.03	18.18		Fail			3.03	27.27
7	Pass	0.82	0.40	54.55	27.27	19	Pass	0.79	0.34	51.52	27.27
	Fail			0.00	18.18		Fail			3.03	18.18
8	Pass	0.76	0.17	45.45	30.30	20	Pass	0.76	0.34	51.52	27.27
	Fail			9.10	15.15		Fail			3.03	18.18
9	Pass	0.48	0.02	30.30	24.24						
	Fail			24.24	21.22						
10	Pass	0.70	0.54	51.52	18.18						
	Fail			3.03	27.27						
11	Pass	0.73	0.48	51.52	21.21						
	Fail			3.03	24.24						
12	Pass	0.67	0.73	54.55	12.12						
	Fail			0.00	33.33						

Table 17(c)

Item Data for Test 6

	Test	p	B	% pass test	% fail test		Test	p	B	% pass test	% fail test
Item						Item					
1	Pass	0.64	0.38	27.27	36.37	13	Pass	0.91	0.13	30.3	60.6
	Fail			3.03	33.33		Fail			0	9.1
2	Pass	0.61	0.57	30.3	30.3	14	Pass	0.55	0.37	24.24	30.3
	Fail			0	39.4		Fail			6.06	39.4
3	Pass	0.67	0.48	30.3	36.36	15	Pass	0.52	0.27	21.21	30.3
	Fail			0	33.34		Fail			9.09	39.4
4	Pass	0.58	0.47	27.27	30.3	16	Pass	0.58	0.28	24.24	36.36
	Fail			3.03	39.4		Fail			6.06	33.34
5	Pass	0.49	0.31	21.21	27.27						
	Fail			9.09	42.43						
6	Pass	0.49	0.45	24.24	24.24						
	Fail			6.06	45.46						
7	Pass	0.52	0.41	24.24	27.27						
	Fail			6.06	42.43						
8	Pass	0.46	0.35	21.21	24.24						
	Fail			9.09	45.46						
9	Pass	0.67	0.33	27.27	39.39						
	Fail			3.03	30.31						
10	Pass	0.24	0.08	9.09	15.15						
	Fail			21.21	54.55						
11	Pass	0.82	0.26	30.3	51.51						
	Fail			0	18.19						
12	Pass	0.49	0.27	21.21	30.3						
	Fail			9.09	39.4						

Table 17(d)

Item Data for Test 7

Item	Test	p	B	% pass test	% fail test	Item	Test	p	B	% pass test	% fail test
1	Pass	0.36	0.63	18.18	18.18	13	Pass	0.79	0.27	21.21	57.57
	Fail			3.03	60.61		Fail			0	21.22
2	Pass	0.7	0.38	21.21	48.48	14	Pass	0.91	0.12	21.21	69.69
	Fail			0	30.31		Fail			0	9.1
3	Pass	0.18	0.49	12.12	6.06	15	Pass	0.82	0.23	21.21	60.6
	Fail			9.09	72.73		Fail			0	18.19
4	Pass	0.36	0.63	18.18	18.18	16	Pass	0.82	0.19	21.21	63.63
	Fail			3.03	60.61		Fail			0	15.16
5	Pass	0.79	0.27	21.21	57.57	17	Pass	0.58	0.14	15.15	45.45
	Fail			0	21.22		Fail			6.06	33.34
6	Pass	0.76	0.13	18.18	57.57	18	Pass	0.36	0.26	12.12	24.24
	Fail			3.03	21.22		Fail			9.09	54.55
7	Pass	0.73	0.35	21.21	69.69	19	Pass	0.55	0.4	18.18	36.36
	Fail			0	9.1		Fail			3.03	42.43
8	Pass	0.91	0.12	21.21	69.69	20	Pass	0.48	0.43	18.18	33.33
	Fail			0	9.1		Fail			3.03	45.46
9	Pass	0.42	0.37	15.15	27.27	21	Pass	0.91	0.12	21.21	69.69
	Fail			6.06	51.52		Fail			0	9.1
10	Pass	0.97	0.04	21.21	75.75	22	Pass	0.59	0.15	21.21	66.66
	Fail			0	3.04		Fail			0	12.13
11	Pass	0.15	0.35	9.09	6.06	23	Pass	0.27	0.52	15.15	15.15
	Fail			12.12	72.73		Fail			6.06	63.64
12	Pass	0.58	0.36	18.18	39.39						
	Fail			3.03	39.4						

The data of that are of interest for diagnostic purpose are (a) the relative proportion of test-masters who pass or fail an item, and (b) the relative proportion of non-test masters who pass or fail an item. Table 17(a) shows the item data for 24

items testing the unit 'Quantum theory of light - The photoelectric effect'. There were 2 items with a large percentage of the group who are test masters but failed the items. For item 4, 33.33 % of the group (75 % of test-masters) are test masters who are misclassified as non-item masters while for item 22, the percentage is 30.31 % (75 % of test masters) . Table 17(b) shows the item data for Test 5 testing the unit ' X-ray - Duality of matter' in which items 9 and 13 resulted in 24.24 % and 51.52 % respectively of the group (or 44 % and 93 % of test-masters respectively) who are test masters but failed these items. Item 10 in Test 6, shown in Table 17(c), testing the unit ' Models of an atom' with 21.21 % of group(or 67 % of test-masters) and item 11 in Test 7, shown in Table 17(d), testing the unit 'Nuclear stability - Radioactivity' with 12.12 % of group (or 60 % of test-masters) were the other items identified where the proportion of test masters who failed the items are very high.

Another matter of concern is when the proportion of the group who are non-masters passing an item exceeds the proportion of the group who are non-masters failing the same item. In Test 4, 100 % of the test non-masters pass items 3, 6 and 11 which are relatively easy items. About six-seventh of test non-masters pass each of the items 4, 5, 8, 10 and 12. About seven-eighth of test non-masters pass item 2 (Test 5), six-seventh pass item 13 (Test 6). For Test 7, seven-eighth of test non-masters pass items 7, 8, 10, 14 and 21.

Figures 14a, b, and c, Figures 15a, b, and c, Figures 16a, b, and c, Figures 17(a), 17(b), and 17(c) show scatterplots of B, r_{pb} and ϕ against difficulty index, p for Tests 4, 5, 6 and 7. Data points inside the top right hand quadrant represent 'good' items that are able to discriminate between 'masters' and 'non-masters' of the tests. Data points outside the quadrant are 'poor' items.

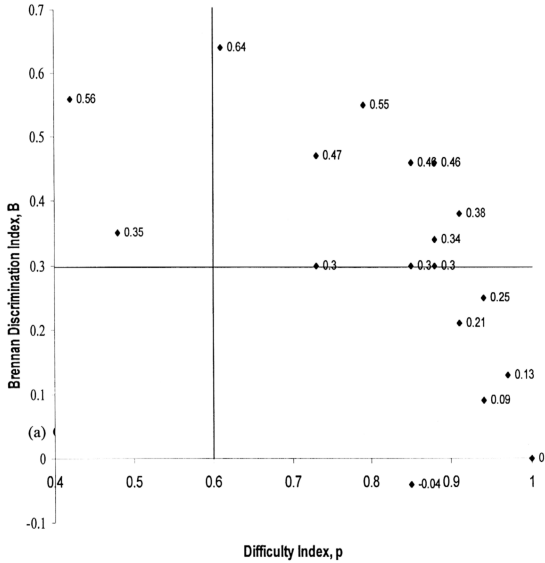


Figure 14(a). Graph of B versus p for Test 4

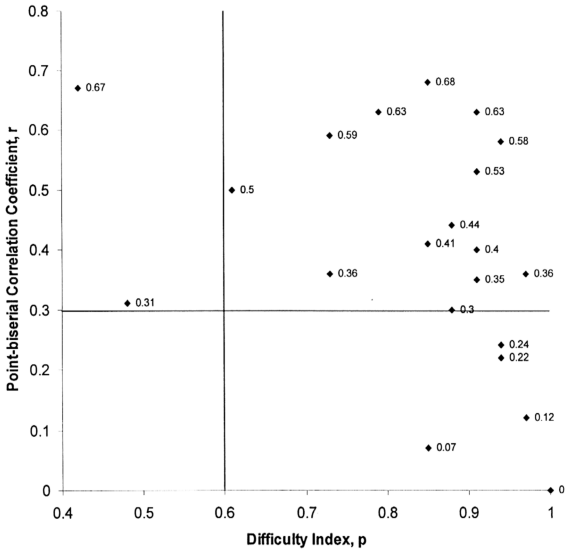


Figure 14(b). Graph of r_{pb} versus p for Test 4

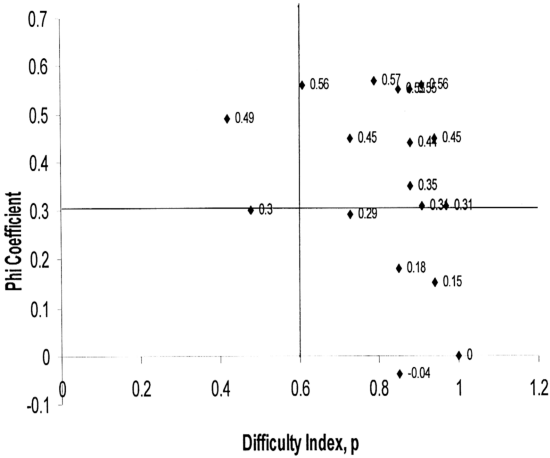


Figure 14(c). Graph of ϕ versus p for Test 4

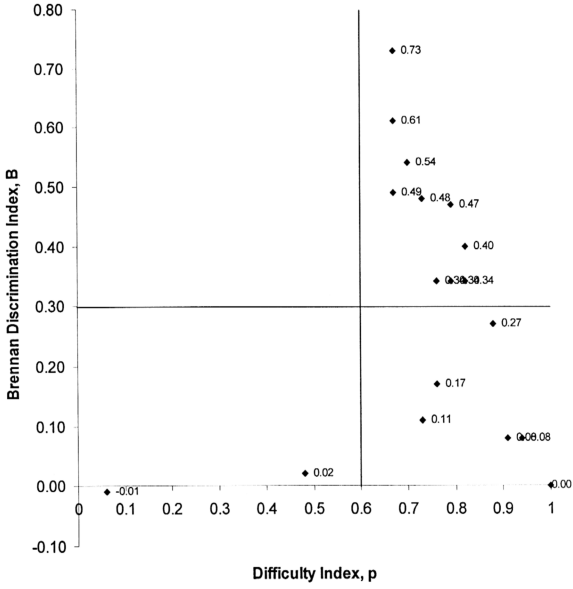


Figure 15(a). Graph of B versus p for Test 5

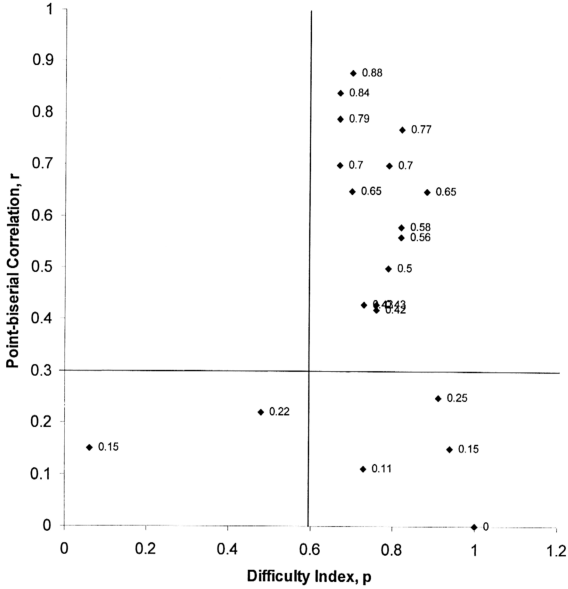


Figure 15(b). Graph of r_{pb} versus p for Test 5

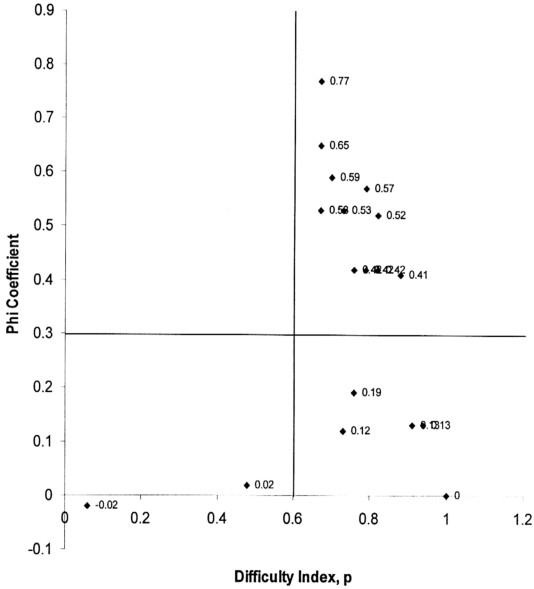


Figure 15(c). Graph of ϕ versus p for Test 5

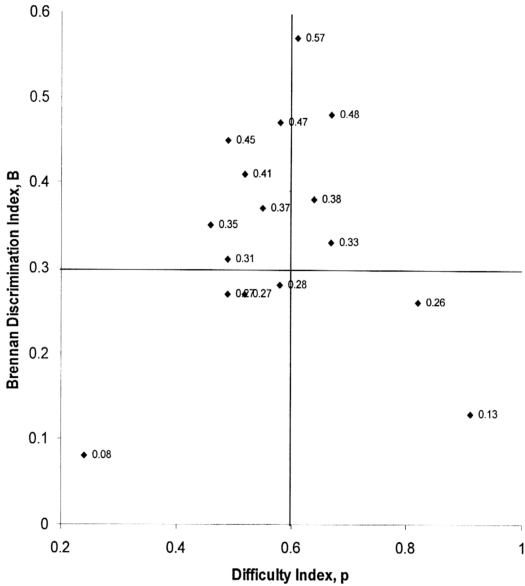


Figure 16(a). Graph of B versus p for Test 6

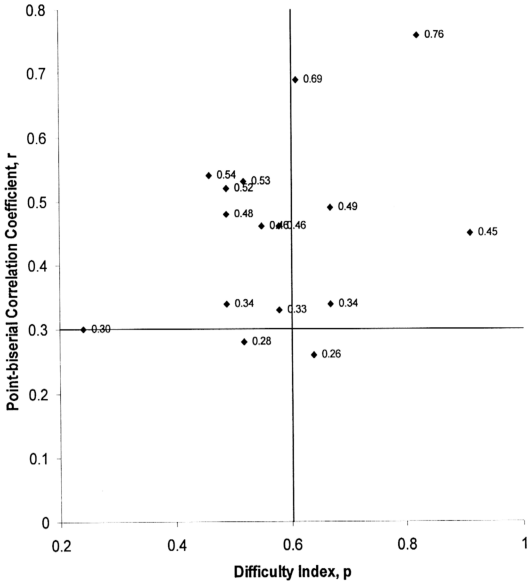


Figure 16(b). Graph of r_{pb} versus p for Test 6

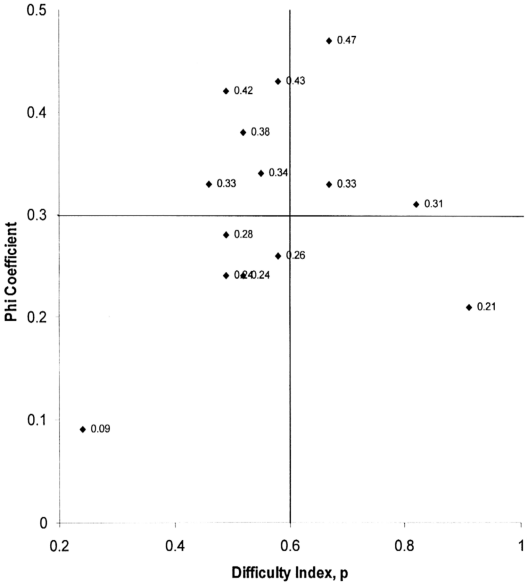


Figure 16(c). Graph of ϕ versus p for Test 6

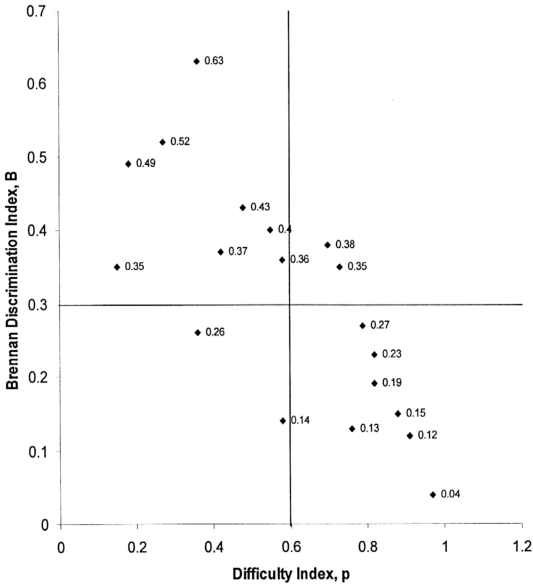


Figure 17(a). Graph of B versus p for Test 7

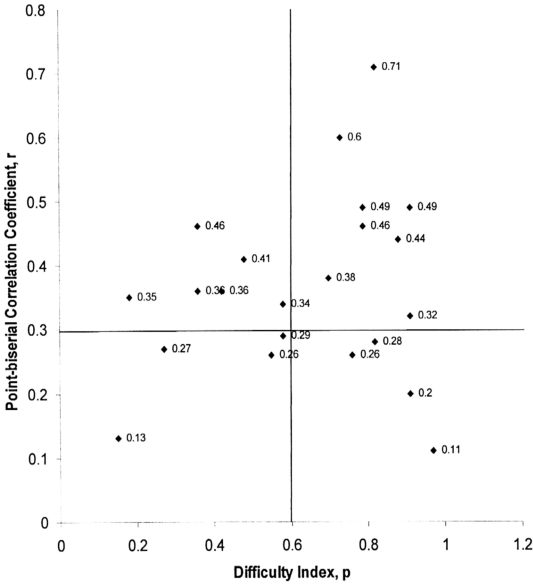


Figure 17(b). Graph of r_{pb} versus p for Test 7

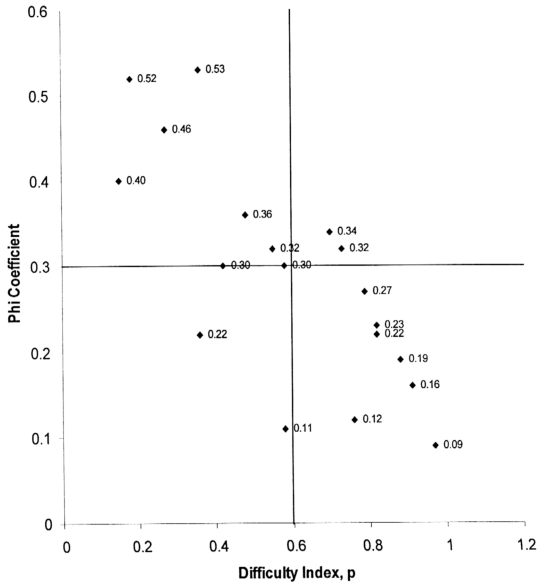


Figure 17(c). Graph of ϕ versus p for Test 7

The 'good' and 'poor' test items for mastery/non-mastery discrimination identified from Figures 14, 15, 16 and 17 are summarised by 2 X 2 contigency tables as shown in Table 18.

Table 18

2 X 2 Contigency Tables Showing the Number of Good and Poor Items

Test 4		*r*		Test 6		*r*	
		good	poor			good	poor
B	good	11	0	B	good	3	1
	poor	5	8		poor	2	10
		r				*r*	
		good	poor			good	poor
phi	good	14	1	phi	good	4	1
	poor	2	7		poor	1	10
		B				B	
		good	poor			good	poor
phi	good	9	6	phi	good	4	1
	poor	2	7		poor	0	11
Test 5		*r*		**Test 7**		*r*	
		good	poor			good	poor
B	good	12	0	B	good	2	0
	poor	2	6		poor	6	15
		r				*r*	
		good	poor			good	poor
phi	good	13	0	phi	good	2	0
	poor	1	6		poor	6	15
		B				B	
		good	poor			good	poor
phi	good	12	1	phi	good	2	0
	poor	0	7		poor	0	21

Table 19 contains a summary of the inter-correlations between the discrimination indices computed using the phi coefficient. This coefficient will indicate whether there is an agreement among the three indices used for mastery and non-mastery decisions.

Table 19

Inter-correlations between Discrimination Indices
Using the Phi Coefficient

	ϕ		
Test	r_{pb} - B	r_{pb} - ϕ	B - ϕ
4	.65*	.73**	.37
5	.80**	.89**	.90**
6	.54*	.71**	.86**
7	.42*	.42*	1.00**

* $p < .05$, 2-tailed.
** $p < .01$

The results indicated that the ϕ coefficients range between .42 to .80 for correlations between r_{pb} and B, .37 to 1.00 for correlations between B and ϕ, and .42 to .89 for correlations between r_{pb} and ϕ. The mean value of ϕ was .60, .69 and .78 respectively for correlations between r_{pb} - B, r_{pb} - ϕ and B - ϕ. The results suggested that the agreement between B and ϕ in identifying 'good' and 'poor' items is the strongest whereas that between r_{pb} and B appears to be the weakest. Generally, the agreement between each pair of indices range from moderately positive to strongly positive.

Table 20 shows the correlations between the item discrimination indices for the multiple-choice tests computed using the Pearson product-moment correlation coefficient r. The difficulty index p for each of the tests is shown in parentheses.

Table 20

Inter-correlations for Discrimination Indices for Tests

Indices	Pearson product-moment correlation, *r*			
	Test 4 (p=.76)	Test 5 (p=.55)	Test 6 (p=.30)	Test 7 (p=.21)
B and r_{pb}	.82	.87	.39	.22
B and ϕ	.87	.99	.97	.97
R_{pb} and ϕ	.87	.90	.50	.18

Figures 18, 19 and 20 show the scatterplots of B and r_{pb}, B and ϕ, r_{pb} and ϕ for the tests.

The data suggested that there is a strong correlation between B and ϕ ranging from 0.87 for Test 4 to .99 for Test 5. These two discrimination indices produce consistent results for the four tests independent of the difficulty levels of the tests. The correlation between r_{pb} and ϕ range from a low of .18 (p= .21) to a high of .90 (p= .55), while that between r_{pb} and B range from .22 (p= .21) to .87 (p = .55). These results appear to indicate a lack of agreement between r_{pb} and B as well as between r_{pb} and ϕ.

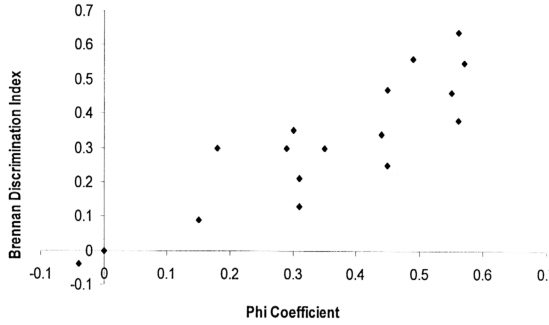


Figure 18(a). Scatterplot of B versus ϕ for Test 4

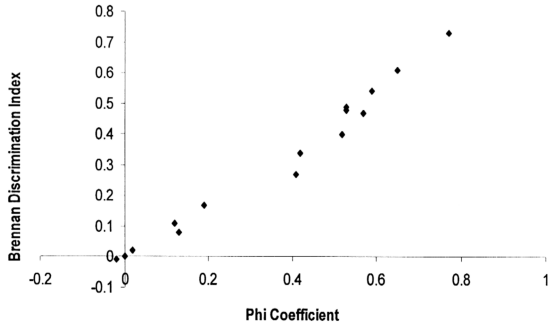


Figure 18(b). Scatterplot of B versus ϕ for Test 5

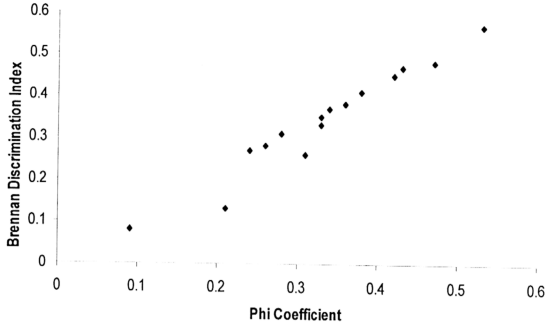


Figure 18(c). Scatterplot of B versus ϕ for Test 6

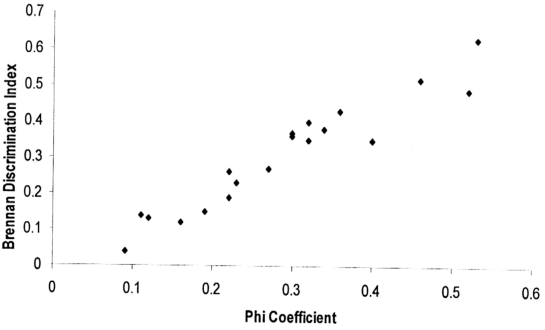


Figure 18(d). Scatterplot of B versus ϕ for Test 7

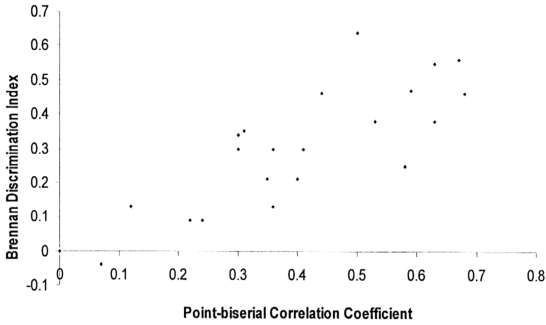


Figure 19(a). Scatterplot of B versus r_{pb} for Test 4

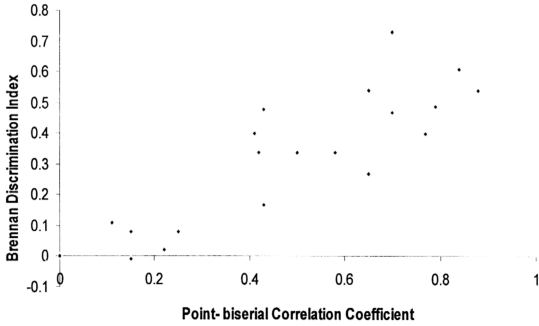


Figure 19(b). Scatterplot of B versus r_{pb} for Test 5

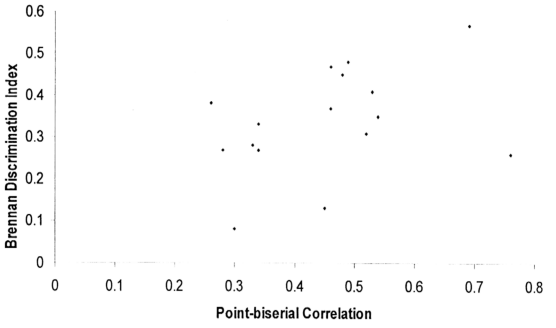


Figure 19(c). Scatterplot of B versus r_{pb} for Test 6

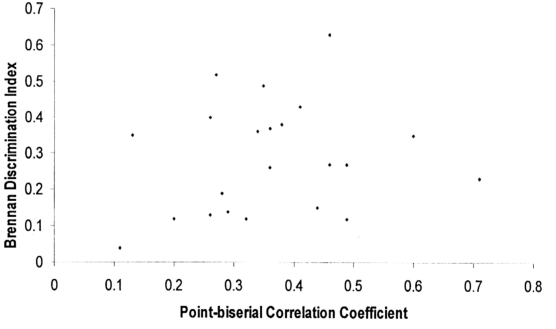


Figure 19(d). Scatterplot of B versus r_{pb} for Test 7

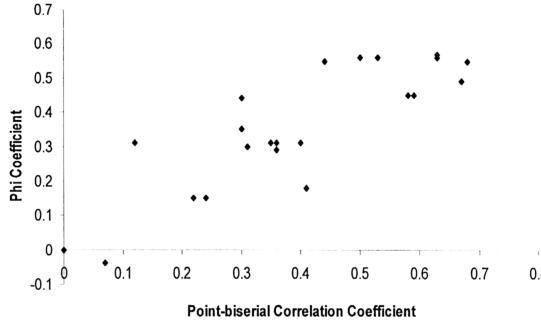


Figure 20(a). Scatterplot of ϕ versus r_{pb} for Test 4

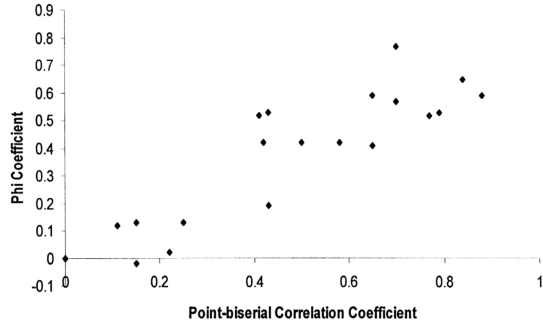


Figure 20(b). Scatterplot of ϕ versus r_{pb} for Test 5

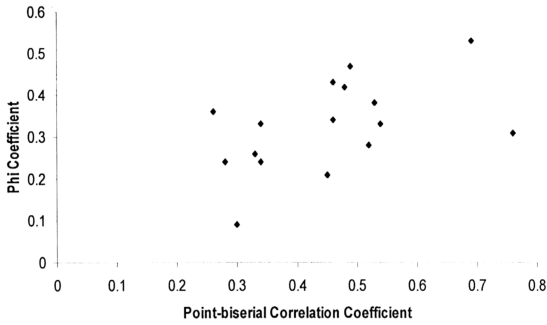


Figure 20(c). Scatterplot of ϕ versus r_{pb} for Test 6

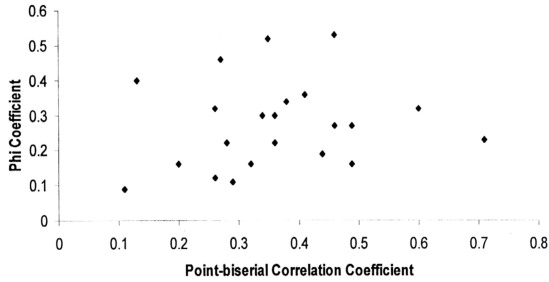


Figure 20(d). Scatterplot of ϕ versus r_{pb} for Test 7