ITEM ANALYSIS OF STUDENT COMPREHENSIVE TEST FOR RESEARCH IN TEACHING BEGINNER STRING ENSEMBLE USING MODEL BASED TEACHING AMONG MUSIC STUDENTS IN PUBLIC UNIVERSITIES

Shafizan Sabri
Sultan Idris Education University
Shafizan72@yahoo.com

ABSTRACT

Beginner string ensemble model based teaching is a systematic instruction which incorporate improvement and new strategies for teaching string ensemble. The instructional design model by Dick and Carey laid the groundwork for these model based teaching. In the design process, achievement test were constructed to evaluate learner's progress and instructional quality. Using student comprehensive test as instrument, the study attempts to examine test items of a researcher made test in the research of string ensemble course for university music students. The quality of each particular item was analyzed in terms of item difficulty, item discrimination and distractor analysis. Reliability test were also conducted in addition to the item analysis to observe the quality of the test as a whole. Two statistical tests were used to compute the reliability of the test: The Kuder-Richardson Formula (KR20) and the Kuder-Richardson Formula 21 (KR21). The comprehensive test was piloted in one university in Perak involving 16 music students. The difficulty index of 41 test items was computed using Microsoft Office Excel. The discrimination analysis was conducted by means of Pearson product-moment correlation using SPSS 17.0. Eventually Microsoft Office Excel was used to compute the reliability statistics. The result indicates that forty four percent of the total test items exceed the difficulty index of 0.8 suggesting easy items. Fifty nine percent of items obtained acceptable range of discrimination index. Distractor analysis reveals that some distractors were not effective. The quality of the item as a whole indicates a reliable value Kuder-Richardson 20 (KR20) value of 0.717 and Kuder-Richardson 21(KR21) value of 0.703. The findings suggest that in order to measure students' performance effectively, necessary improvement need to be done where items with poor discrimination index should be reviewed.
Keywords: Item analysis, difficulty index, discrimination index, distractor analysis

INTRODUCTION

The instructional design of model based teaching for experimental group was design with the basis of framework by Dick and Carey model. In order to test the effectiveness of the designed model based teaching, the researcher applied quantitative approaches. One of the main research instruments involves a researcher made test to be exact comprehensive test of string instruments.

The comprehensive test for research in teaching beginner string ensemble may perhaps be classified as norm-referenced tests. Boyle and Radocy (1987) distinctly indicate that the main function of a norm-referenced test is to distinguish the performance amongst student in a particular test. In addition, Ary, Jacobs and Razavieh (2002) denote that in norm-referenced tests, individual's achievement is elucidated with reference to the performance of others within the specified group of interest. Fraenkel and Wallen (2008) and Ary *et al* (2002) agree that norm-referenced tests bestow researchers with a broad range of scores. Thus, the difficulty of the item and the discriminatory power of each test items should be a matter of concern. Matlock-Hetzel (1997), in her study of basic concepts in item and test analysis apparently advocates that conducting item and test analyses is an essential step in evaluating the efficacy of educational programs or courses whereby norm-referenced tests is constructed for the usage in instructional purposes.

Instructional Assessment Resources (IAR 2011) believes that "an item analysis involves many statistics that can provide useful information for improving the quality and accuracy of multiple-choice or true/false (question)". The quality of each particular item was analyzed to evaluate the quality of each item in terms of item difficulty and item discrimination. Item difficulty is basically the proportion of students who responded correctly to an item. In the meantime, item discrimination is a measure to differentiate between the performance of students in the high score group and those in the low score group. In addition to the above measures, distractor analyses were conducted to determine comparative efficacy of each distractors in the multiple choice items. Thompson and Levitov in Matlock-Hetzel proposed that the quality of the test as a whole can be evaluated by means of computing reliability.

Using student comprehensive test as instrument, the purpose of the study, therefore, attempts to examine test items of a researcher made test in the research of string ensemble course for university music students.

METHODOLOGY

Participants

The pilot test was conducted at one university in Perak. The subjects consist of diploma students from the beginning string ensemble class enrolled for that particular semester. The students were from the beginner string ensemble. The class consists of sixteen students with various music backgrounds. Their strings instruments were selected before commencing the class. The researcher did not have any involvement in the string selections process. The class was conducted with eleven violin, two viola, two cello and two double bass students. In this trial, the classroom was set up in the string ensemble setting. During each sessions, researcher used LCD projector, video camera, mixer, microphone, YouTube material, exercises handout and handy-cam. The researcher used the two selected classroom which was the recital room and selected string's room. The students have to follow the 14 weeks two hour class schedule on every Tuesday which has been schedule by the academic department. At the end of the 14 week class, the participants completed a post-test. The comprehensive string instruments test took place during the timetabled class hours. Participants were given 45 minutes to complete the test. All the test booklets were collected and kept well for

data analysis procedure. The test papers were then rated by the researcher and student's scores were calculated.

Materials

The string instruments comprehensive test was constructed to evaluate the cognitive aspect of the students focusing on what the person knows (Schumacher and Mcmillan, 2006). In addition, they also clarify that "the purpose of achievement tests is to measure what has been learned, rather than to predict future performance" (p.191).

The comprehensive test was based on beginning string ensemble contents. The test consists of multiple choice and short answer question and was modelled after the Bloom taxonomy which identifies six major categories of objectives specifically knowledge, comprehension, application, analysis, synthesis and evaluation. These comprehensive tests integrate three objectives to be exact knowledge, comprehension and application. Section A comprises 15 multiple choice items for testing specific knowledge and application in string instrument. Each question was followed by a set of alternate answers. Each set contains four-choice items specifically A, B, C and D with one correct answer. Section B and C comprises two and five short item question each.

Data Collection and Scoring of Comprehensive Test

Data was assembled from sixteen examinee's answer to items in the comprehensive test for beginner string ensemble level of university student. A total of 41 items were utilized in the item analysis process. In section A which comprises 15 items, each correct answer was given one point and zero for each wrong answer. In section B, each correct answer was awarded one point and zero each wrong answer. Section C includes five multipoint items types of question. Based on Boyle and Radocy guidelines for multipoint item discrimination index, items for section were analyzed for discrimination index as follows. For question three, the entire item is regarded totally correct if the student managed to list all the correct answer to the question. If any one of the answers is incorrect or missing then the answer is regarded as wrong. Question four requires students to correctly notate the fingering scale for instruments. The item is regarded totally correct if the student managed to notate all the correct answer. Similarly, if any one of the answers is missing or incorrect, the answer is regarded as wrong. For question five, the entire item is regarded totally correct if the student managed to list four ways of caring and maintaining string instruments. If any one of the answers is incorrect or missing then the answer is regarded wrong.

Item Statistics

Item Difficulty

The results of students' achievement in this comprehensive test were then utilized to determine the quality of each particular item in terms of item facility, item discrimination and distractor analysis. Item difficulty, commonly known as *p*-value refers to the proportion of examinees that responded to the item correctly. The *p*-value is calculated using the following formula:

$p = R / T$

where p = item difficulty index

R = the number of correct responses to the test item

T = the total number of responses comprises both correct and incorrect responses

The p-value ranges from 0.0 to 1.00. A high p-value indicates an easy item. Instructional Assessment Resources (IAR) acknowledged values of difficulty index and their evaluation as tabulated in Table 1.

Table 1: Evaluation of Item Difficulty for Item Analysis

Item Difficulty Index (p)	Item Evaluation
Above 0.90	Very easy item
0.62	Ideal value
Below 0.20	Very difficult item

Source: Instructional Assessment Resources (IAR 2011)

Item Discrimination

Matlock–Hetzel clarified two ways of determining discriminating power of test item, to be exact the Discrimination Index and Discrimination Coefficient. Matlock-Hetzel, Si-Mui and Rasiah (2006) Mitra, Nagaraja, Ponnudurai and Judson (2009) and Boopathiraj and Chellamani (2013) defines item discrimination as a measure used to discriminate between students in the top with that of the low group who obtained the correct responses. Fundamentally, the discrimination index differentiates students who are knowledgeable and those who are not, meticulously revealing top scorers and low scorers achievement in each item. The value of discrimination index ranges between -1.0 to 1.0.

Item discrimination index (D) is calculated by the following formula D = (UG-LG)/n. Where D = discrimination index, UG = the number of students in the upper group 27% who responded correctly, LG = the number of students in the lower group 27% who responded correctly and

n = number of students in the upper or lower group. The items were classified accordingly to their discrimination index with reference to Ebel's (1972) guidelines.

Table 2: Evaluation of Discrimination Indexes for Item Analysis

Index of Discrimination	Item Evaluation
0.40 and above	Very good items; accept
0.30 – 0.39	Reasonably good but subject to improvement
0.20 – 0.29	Marginal items usually need and subject to improvement
Below 0.19	Poor items to be rejected or improved by revision

Source: Ebel (1972) in Ovwigho (2013)

Matlock-Hetzel emphasized the advantage of using discrimination coefficient instead of discrimination index. Discrimination coefficients includes every single person taking the test despite the fact that only the upper (27%) and lower scorer (27%) are included in the discrimination index calculation process. According to Instructional Assessment Resources (IAR), Le (2012), Owigho , and El-Uri and Malas (2013), discrimination coefficients is a measure using point biserial correlation. The correlation, commonly known as Pearson product-moment correlation is computed

to determine the relationship between student's performance in each item and their overall exam scores. This paper plump for utilization of discrimination coefficient considering the small sample size of 16 students with the intention that every single person performance was taken into consideration. The discrimination coefficient, the Pearson *r*, for each item was computed using Statistical Package for the Social Sciences (SPSS) version 17.The Pearson, *r* coefficient ranges between -1 and 1. Parallel to the discrimination index, a higher value indicates a powerful discrimination power of the respective test. A highly discriminating item reveals that students with high score got the item right and students with low score answer the item incorrectly. Items with negative values should be rejected for the reason that negative value reflects the opposite effects of discriminating power for that particular item.

Distractor Analysis

Distractors are classified as the incorrect answer in a multiple-choice question. According to Instructional Assessment Resources (IAR), student performance in an exam item are very much influence by the quality of the given distractors. Hence, it is necessary to determine the effectiveness of each item provided distractor as an addition to the item difficulty analysis. Analysis was conducted only for items in section A of the comprehensive test since distractor analysis are associated only with multiple-choice formats.

Statistical Analysis

The data from the item difficulty and item discrimination analysis were each conveyed as mean and standard deviation of the total number of items. SPSS 17 were employed in verifying the relationship between the item difficulty index and discrimination coefficient for each test item using Pearson correlation.

Test Statistic

The data of total score for each student were statistically analyzed in terms of mean, median and standard deviation of the total number of students.

Reliability

Reliability is expressed as the constancy of particular instruments in producing the same result in repeated measurements. An instrument is considered reliable if the instrument produce same result every time when use to evaluate identical measurement. Boyle and Radocy proposed using Kuder Richardson formula for analyzing test with dichotomous items. Data from string instruments were divided into two sections. Kuder-Richardson 20, a formula which is based on item difficulty was used to analyse internal consistency of section A in the string instrument comprehensive test. The value of KR20 range between 0 to 1.The closer the value to 1 the better the internal consistency. The KR20 formula is commonly used to measure the reliability of achievement test with dichotomous choices. According to Fraenkel and Wallen, one should attempt to generate a KR20 reliability coefficient of .70 and above to acquire reliable score. The formula for estimating reliability is as follows:

$$KR20 = \frac{n}{n-1}\left(\frac{SD^2 - \sum PQ}{SD^2}\right)$$

Where n = number of items

SD^2 = variance of scores on the test (square of the SD (standard deviation))

P = proportion of those who responded correctly

Q proportion of those who responded incorrectly

Source: Wiseman (1999, p. 102)

The total reliability of comprehensive test (section B and C) was analyzed using KR21 formula. Unlike KR20, KR21 formula assumed that all items comprise equal difficulty. The formula for calculating KR21 is as follows:

$$KR21 = \frac{n}{n-1}\left(1 - \frac{M - \frac{(M)^2}{n}}{SD^2}\right)$$

Where n = number of items on the test

M = mean of the scores on the test

SD^2 = variances of scores on the test [square of the SD (standard deviation)]

Source: Wiseman (p. 103)

RESULTS

Item Statistic

Item Statistic was employed to evaluate the performance of individual test items utilizing student's responses to each test items.

Item Difficulty

The indices of item difficulty level of each item are presented in Table 7.

Table 7: Item Difficulty of Comprehensive Test Items

Item No.	Item Difficulty	Item No.	Item Difficulty	Item No.	Item Difficulty
1	0.44	16	0.44	31	1.0
2	0.44	17	0.44	32	0.94
3	0.81	18	0.25	33	0.75
4	0.69	19	0.63	34	0.50
5	0.75	20	0.38	35	0.94
6	0.69	21	0.81	36	0.94
7	0.75	22	0.56	37	0.88
8	0.81	23	0.13	38	0.81
9	0.56	24	0.44	39	0.81
10	0.88	25	0.31	40	0.56
11	0.88	26	0.94	41	0.38
12	0.50	27	0.88		
13	0.81	28	0.94		
14	0.56	29	1.0		
15	0.38	30	1.0		

Based on recommendations by Instructional Assessment Resources (IAR), test items were classified into three categories in terms of level of difficulty as indicated in Table 8.

Table 8: Distribution of Items in Terms of Level of Difficulty in Categories

Item Difficulty Index (*p*)	Total Item
Easy (Above 0.90)	8
Moderate (0.20 – 0.90)	32
Difficult (Below 0.20)	1

Item Discrimination

The coefficient of item discrimination each item is indicated in Table 9.

Table 9: Comprehensive Test Item Discriminations in terms of Discrimination Coefficient

Item No.	Discrimination Coefficients	Item No.	Discrimination Coefficient
1	0.627	22	0.624
2	-0.424	23	0.596
3	0.244	24	0.752
4	0.219	25	0.487
5	0.580	26	0.029
6	0.236	27	0.230
7	0.150	28	-0.022
8	0.149	29	0
9	0.374	30	0
10	0.305	31	0
11	0.455	32	0.183
12	0.534	33	0.294
13	0.403	34	0.558
14	0.149	35	0.234
15	0.093	36	-0.074
16	0.627	37	0.155
17	0.527	38	-0.105
18	0.767	39	-0.010
19	0.625	40	-0.102
20	0.606	41	0.067
21	0.308		

Table 8 reveals classifications of test items into five level of discrimination in terms of discrimination coefficient based on recommendations by Ebel (1972) in Ovwigho (2013).

Table 8: Distribution of Items in Terms of Level of Discrimination in Categories

Discrimination Coefficient	Total Item
Very Good (above 0.40)	15
Reasonably Good (0.30 – 0.39)	3
Marginal (0.20-0.29)	6
Poor (Below 0.19)	17

Distractor Analysis

The distractor analysis conducted on section A of student's comprehensive test yields the following results tabulated in Table 9. Zero responses designates the distractor was not selected by any of the students.

Table 9: Response Frequency Distribution of Items in Comprehensive Test

Item	Scorers	A	B	C	D	Item	Scorers	A	B	C	D
1	Top	2	1	4*	1	9	Top	2	1	4*	1
	Low	3	1	3*	1		Low	2	1	5*	0
2	Top	2	3	2	1*	10	Top	0	0	1	7*
	Low	0	2	0	6*		Low	0	0	1	7*
3	Top	0	0	2	6*	11	Top	0	7*	1	0
	Low	0	0	1	7*		Low	0	7*	0	1
4	Top	3	0	0	5*	12	Top	1	0	3*	4
	Low	2	0	0	6*		Low	0	1	4*	3
5	Top	0	3	0*	5	13	Top	0	1	6*	1
	Low	0	1	7*	0		Low	0	0	7*	1
6	Top	4*	3	1	0	14	Top	6*	1	1	0
	Low	7*	0	0	1		Low	3*	3	2	0
7	Top	0	2	0	6*	15	Top	6	1	1*	0
	Low	0	2	0	6*		Low	3	0	5*	0
8	Top	6*	2	0	0						
	Low	7*	1	0	0						

Items mark with * indicates the correct answers

A distractor analysis assists in distinguishing plausible distractors from implausible ones. A total of 30 distractors were regarded as implausible due to the fact that those distractors were selected neither by the top scorer nor the low score. Item 15 clearly indicates a confusing item seeing that distractor A is selected by more student than the correct answer C.

Statistical Analysis

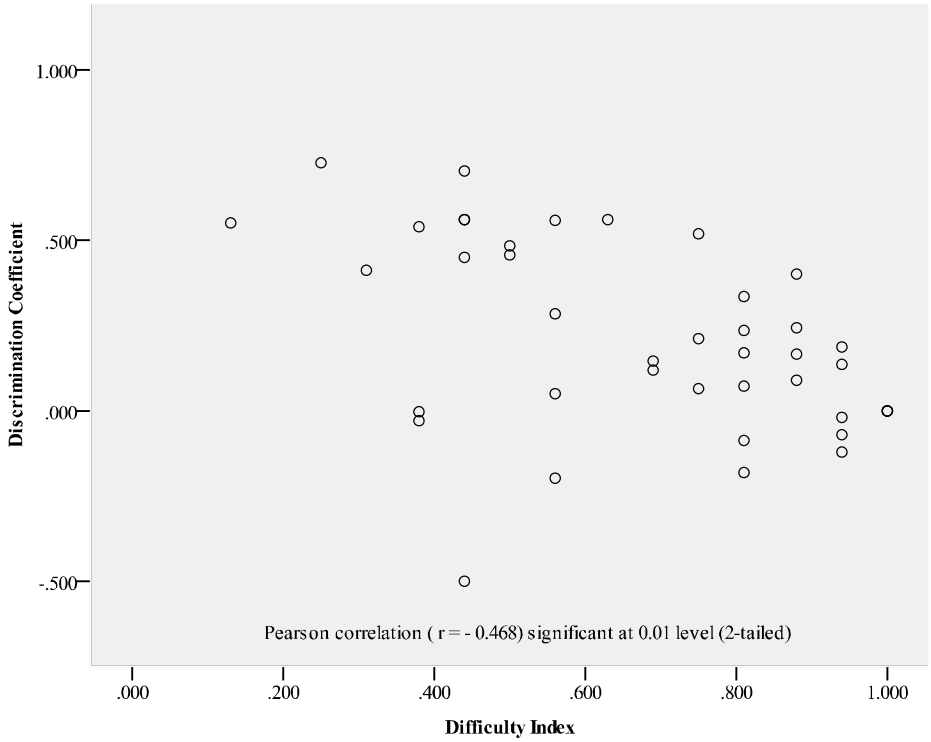
Table 10 provides the results obtained from the analysis of the item difficulty index and item discrimination coefficient.

Table 10: **Descriptive Statistics for item difficulty index and item discrimination coefficient**

	N	*M*	*SD*
Item Difficulty Index	41	0.673	0.236
Item Discrimination Coefficient	41	0.214	0.278

Figure 1 provides scatter plot showing relationship between difficulty index and discrimination coefficient of items. Also revealed is the Pearson correlation value, computed between the two variables. It can be concluded that there is a significant week negative correlation between the two variables ($r = -0.468$, $p = 0.02$) with increasing value of difficulty index, there is decrease in discrimination coefficient.

Figure1: Scatter Plot Showing Relationship between Difficulty Index and Discrimination Coefficient of Items



Test Statistic

Descriptive Statistic

The data obtained from the comprehensive test were used in the statistical analysis. Students' scores are organized in a frequency distribution as shown in Table 3. Table 4 provides the results obtained from the analysis of student comprehensive test score.

Table 3: Frequency Distribution of Score for 16 Music Students on Beginner String Ensemble

Score	Frequency
94	1
88	1
82	1
80	1
78	1
72	1
70	1
68	1
66	1
64	1
62	3
60	3
Total	**1128** **16**

TABLE 4: Descriptive Statistics for Comprehensive Test Score

	Descriptive Statistic
N (total number of student)	16
Mean	70.5
Median	67
Standard Deviation	10.820

Reliability Coefficient

Reliability coefficient was employed to evaluate the performance of the test as a whole. Result of string instrument comprehensive test for section A is shown in table 5. The computed KR20 of the comprehensive test (section A) is .717. According to Fraenkel and Wallen, one should attempt to generate a KR20 reliability coefficient of .70 and above to acquire reliable score. This value of KR20 appears to be reliable thus revealing that this comprehensive test is a reasonably reliable instrument. The total reliability of comprehensive test (section B and C) was analyzed using KR21 formula. Results of students achievement in section B and C is shown in table 6.

TABLE 5
String instruments comprehensive test (Section A) Scores

Students	Scores
1	13
2	11
3	9
4	11
5	11
6	5
7	6
8	6
9	7
10	6
11	12
12	12
13	11
14	15
15	13
16	12

TABLE 6
Result of string instruments comprehensive test (section B and C)

Students	Scores
1	25
2	30
3	28
4	34
5	22
6	24
7	31
8	18
9	25
10	25
11	24
12	20
13	20
14	23
15	23
16	33

Microsoft excel was used to analyze the statistical data for section B and C. Result from the analysis show that the KR21 for the comprehensive test (section B and C) is .703. The reliability coefficient for comprehensive test (section B and C) appears to be satisfactory with reference to Fraenkel and Wallen recommendation stated earlier. This indicates that this test is reasonably reliable instrument.

DISCUSSION

Student comprehensive test was constructed to evaluate the cognitive aspect of the students in the beginner string ensemble. The item analysis conducted in this study implicates three statistics to assists in analyzing the effectiveness of each of the comprehensive test questions specifically item difficulty, item discrimination and distractor analysis.

Item difficulty lends a hand in distinguishing easy item from difficult ones. In general, there is a good distribution of difficulty throughout the test. Analysis of item difficulty on 41 items of the comprehensive test denotes that 78% of the test items were in the moderate level of difficulty. Only 44% of the total test item possess difficulty index of over 0.8. Mitra *et. al* (2009) reported that 40% of the multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests had the difficulty level over 0.8. In the meantime, Si Mui and Rasiah in a study analyzing year two examinations of a medical school found that 40% of the multiple choice question (MCQ) surpassed the difficulty level of 0.7. 20% of the items with difficulty level of 0.2 and over were classified as easy items with three questions acquires difficulty index of 1.0 and only 2% were determined to be difficult questions. Instructional Assessments Resources (IAR) insinuates the usage of easy question as warm up questions in assessing student mastery. The items that were classified as difficult items ought to be reconsider in terms of language and content appropriateness. A low value of difficulty

index may possibly indicate a miskeyed item. Additionally, it may also indicate that the tested topic were inappropriate In spite of this, difficult questions which results in frustration for some students function as challenge among the top student as evokes by Schreyer Institute for Teaching Excellence.

Reviews of literature on the subject of item analysis disclose numerous methods for evaluating the discriminatory power of individual items for instance discrimination index, biserial correlation coefficient, point biserial coefficient and phi coefficient. In this study, the point-biserial coefficient, typically known as Pearson product-moment correlation suggested by Matlock-Hetzel were selected in calculating the discrimination power. In general a total of 24 items (59% of the total comprehensive test items) can be categorized as acceptable with point-biserial figures of over 0.20. Inspection on results of the item discrimination analysis discloses a good number of items with very good discriminating power. 37% of the items regarded as very good items, accomplished Pearson correlation, *r* of over 0.4. Even though the overall discriminatory accomplishments of the comprehensive test were satisfactory, some crucial caveats ought to be considered. A total of 41% (17) of the total items were classified as poor discriminating items (less than 0.20) with three items (18%) did not discriminate at all and six items (35%) with negative discrimination coefficient. This finding was similar to a study conducted by El-Uri and Malas to analyze undergraduate examination in obstetrics and gynaecology, who reported that 38% of the test items had the discrimination coefficient less than 0.2 with 23 questions obtained negative discrimination. Items with poor and negative discrimination coefficient should be highlighted for reviewing purpose. A poor discriminating power might signify confusing items which were ambiguously worded or indicates a miskeyed item. Ultimately, items with negative coefficient should be removed from the comprehensive test. Si Mui & Rasiah and Matloct-Hetzel coincide in the reasoning of the negative value. They proposed that student in the low achievement group often make a guess in answering the easy question and by chance come up with the correct answer. Contradictory, students in the upper achievement group embark upon the easy question too vigilantly and end up choosing the wrong answer. Items with negative discrimination coefficient should be eliminated from the test as put forward by El-Uri and Malas and Ovwigho . The reason is that item with negative discrimination coefficient indicates students with low score got the item right and students with high score answer the item incorrectly.

A distractor analysis assists in distinguishing plausible distractors from implausible ones. A high percentage of 70% from the total distractors were regarded as implausible due to the fact that those distractors were selected neither by the top scorer nor the low score. One item clearly indicates a confusing item seeing that one distractor is selected by more student than the correct answer.

Analysis of difficulty index together with discrimination coefficient reveals a total of 41% of the test items with poor discriminating index had the difficulty index ranging from 0.38 to 0.94. 59% of the 41 test items with acceptable discrimination index had the difficulty index ranging from 0.13 to 0.94. 15 out of 41 items with very good discrimination coefficient acquire the lowest difficulty index of 0.13 and the highest difficulty index of 0.88. When difficulty index and discriminating coefficient were put side by side, it is noticeably that item with similar level of

difficulty possess diverse discrimination coefficients. Si Mui & Rasiah denotes this divergence as a result of students who makes a guess when selecting the correct responses.

In the process of establishing the quality of the test as a whole, the formula developed by Kuder and Richardson (Wiseman) were used to compute the reliability of the test: The Kuder-Richardson Formula (KR20) and the Kuder-Richardson Formula 21 (KR21). The computed KR20 of 0.717 and KR21 of 0.703 of the comprehensive test which appear to be satisfactory indicate that this test is a reasonably reliable instrument in producing consistent scores. Similarly, Lin, Tseng and Wu (1999) while doing item analysis on multiple choice test items in the Registered Nurse Licensure Exam reported a KR20 value range 0.86 to 0.94 in the analysis of internal consistency.

Boyle and Radocy in their book highlighted the importance of conducting item analysis. They advocate that item analysis facilitates test developer in the process of test quality enhancement which typically involves assessment cycles of preserve, eliminate, improve or endorse particular item. Problematic items specifically items with ambiguous wording and wrongly keyed be reviewed based on the calculated difficulty index and discrimination coefficient values to improve the quality of the test. Content expert should be consulted to improve items identified as problematic in terms of language and content appropriateness.

CONCLUSION

Results of the study disclose that comprehensive test items with good discrimination coefficient have a tendency to befall in the range from difficult item to easy item. Alternatively, items with negative discrimination demonstrate difficulty ranging from moderately difficult to very difficult range. Auxiliary inspection of the extensive spread of discrimination is indispensible prior to removing items with poor or negative discrimination. Factors which instigate such poor discrimination should be delicately considered. Item analysis alleviates test developer in developing an ideal achievement test which functions as tools to evaluate learner's progress and instructional quality in the model based teaching for teaching beginner string ensemble among students in public universities.

REFERENCES

Ary, D., Jacobs L.C. & Razavieh, A. (2002). Introduction to Research in Education. (6th ed.). California: Wadsworth.

Boopathiraj, C. & Chellamani, K. Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in Education. International Journal of Social Science and Interdisciplinary Research 2013; 2:2

Boyle, J.D & Radocy, R.E. (1987). Measurement and Evaluation of Musical Experiences. NewYork: Macmillan.

El-Uri, F.I & Malas, N. (2013). Analysis of Use of A Single Best Answer Format in An Undergraduate Medical Examination. *Qatar Medical Journal* 2013:1.

Fautley,M. (2010). Assessment in Music Education. New York: Oxford University Press.

Fraenkel, J.R. & Wallen, N.E. (2008). *How to Design and Evaluate Research in Education* (7th ed.). New York: McGraw-Hill.

Institute for Interactive Media and Learning (2013). *Multiple Choice Questions*. Retrieved November 30, 2013, from University Of Technology Sydney, Institute for Interactive Media and Learning Web site: http://www.iml.uts.edu.au/assessment/types/mcq/index.html

Instructional Assessment Resources. (2011*). Item Analysis*. Retrieved November 9, 2013 from University of Texas at Austin, Instructional Assessment Resources, IAR Web site: http://www.utexas.edu/academic/ctl/assessment/iar/students/report/itemanalysis.php

Le, L.T. (2012). *Item Point-biserial Discrimination*. Retrieved November 29, 2013 from Australian Council of Educational Research Web site: http://www.acer.edu.au/documents/Conquest-Notes-5-ItemPointBiserialDiscrimination.pdf

Lin, L.C, Tseng, H.M & Wu, S.C. (1999). Item Analysis of the Registered Nurse Licensure Exam Taken by Nurse Candidates from Vocational Nursing High Schools in Taiwan. Proc Natl Sci Counc 1999; 9(1): 24-31.

Matlock-Hetzel, S. (1997 January). *Basic Concepts in Item and Test Analysis*. Paper Presented at the Annual Meeting of the Southwest Educational Research Association, Austin.

Mcmillan, J.H & Schumacher, S. (2005). *Research in Education: Evidence-Based Inquiry* (6th ed.). London: Pearson.

Mitra, N.K, Nagaraja, H.S, Ponnudurai, G. & Judson, J. P. The Levels of Difficulty and Discrimination Indices in Type A Multiple Choice Questions of Pre-clinical Semester 1 Multidisciplinary Summative Tests. E-Journal of Science, Medicine and Education 2009; 3(1):2-7.

Ovwigho, B.O. Empirical Demonstration of Techniques for Computing the Discrimination Power of a Dichotomous Item Response Test. IOSR Journal of Research and Method in Education 2013; 3(2): 12-17

Schreyer Institute for Teaching Excellence. (2013). *Improve Multiple Choice Tests Using Item Analysis*.Retrieved November 28, 2013, from Pennsylvania State University, Schreyer Institute for Teaching Excellence Web site: http://www.schreyerinstitute.psu.edu/Tools/ItemAnalysis/

Sim, S.M & Rasiah R.I. Relationship between Item Difficulty and Discrimination Indices in True/False Type Multiple Choice Questions of a Para-clinical Multidisciplinary Paper. Ann Acad Med Singapore 2006; 35: 67-71

Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities. Article. Full-text available.Â Rasch model analyses adduced a fitting item pool, after the deletion of 39 items. The resulting item difficulty parameters were used for the comparison of the different formats. The multiple choice format â€œ1 of 6â€ differs significantly from â€œx of 5â€, with a relative effect of 1.63, while the multiple choice format â€œx of 5â€ does not significantly differ from the free response format.