Analysis of Difficulty and Point- Biserial Correlation Indices of 2014 Akwa Ibom State Mock Multiple Choice Mathematics Test

Cyrinus B. Essen, Ph. D
Department of Educational Foundations, Guidance & Counselling
Faculty of Education
University of Calabar,
Calabar, Cross River State
ecyrinus@ymail.com, cyrinusessen@gmail.com

Godwin Akpan, Ph. D
Department of Educational Foundations
College of Education, Afaha Nsit,
Akwa Ibom State

Abstract
The study was carried out to examine the difficulty, point-biserial correlation indices and the extent of the relationship in 2014 Akwa Ibom State Mock 50 items multiple choice mathematics test. Three research questions were raised that guided the study. Expo facto research design was adopted for the study. The sample consisted of 3,094 examinees responses to the multiple choice mathematics randomly selected from a population of 47, 600 senior secondary two (SS2) responses after they were duly scored by the markers. The responses were subjected to Pearson Product Moment Correlation statistics using IBM SPSS statistics version 20 to examine the quality of the test items, assessing difficulty indices, the point biserial correlation coefficient indices and the relationship between them. The results revealed that 39 items were considered as being moderately difficult with acceptable p-value range of .71 - .74. Eleven items were considered as being easy/poor with p-value range of .75 -.92. Ten items had point –biserial correlation indices of .20 marginally acceptable value, while, 40 items had point –biserial correlation coefficients of .19 and below classified as displaying poor discrimination power. The relationship between difficulty and point-biserial correlation coefficient indices were not linear. It was suggested that there is need to carry a proper item analysis of the 2014 multiple choice mathematics items to ensure effective reliability and validity of the items; an outright removal of the easy/poor items and those with poor discrimination values, for future use and dependable item bank.

Key words: *Assessing difficulty, Point- biserial correlation, Indices, Mock multiple – choice, Mathematics test.*

Introduction

The analysis of the contributions of each item to quality of test administered for decision-making in the field of education is imperative, for drawing an inference about the examinees' abilities, the skills measured, the reliability and validity of such measuring instrument. Test item analysis, provides reliable information about the state of each question or item that makes the test and a source of assessment to the teachers and the test developers. The inclusion of problematic items inadvertently, makes the inference drawn about the examinees' ability and the use of such instrument in decision making illusive (Verma, 2008).

Verma (2008) and Shakil (2008) explained that, both qualitative and statistical item analyses are required if the instructors and test developers are to be guided for the improvement of instruction. Qualitative item analysis is often carried out by content experts and test review boards in identifying items that do not appear to meet minimum quality control criteria. This quality review is done during item development. While statistical item analysis is conducted after test had been administered and the real-world data or students' responses are gathered. Statistical analysis of dichotomously ("1" correct and "0" incorrect) scored test items ensure that items effectively evaluate examinee learning and discriminate between those that master the subject from those that do not. These analyses are essential if quality of items and examinees' abilities and proficiency are the prime factors in quality assessment in education.

The identification of problematic items sometimes known as bad or misfit items could distort examinees responses, thereby resulting in wrong decision making in the use of the test results, either for promotional, employment or scholarship purposes. Thompson (2009) stressed that the goal of item analysis is to use detailed statistics to determine possible flaws in the item. According to Brown and Frederick (1971) cited in Rana (2014), item analysis helps to identify defective test items that the learners do not master the contents. Thus, the effectiveness of individual items in terms of difficulty levels and the power to discriminate between the high and low scorers is ensured. This allows for poor items to be rejected and good items retained for future use. Besides, Krishnan (2013) stressed that particular attention should be given to individual items, item characteristics, the probability of answering items correctly, the overall ability of the test taker, and the extent to which an item conforms with the rest of the items in a test.

Two complementary theories in psychometrics provide fundamental guides for items analysis in the field of education. These are: Classical test theory and Item response theory. However, this paper is concerned with using two major statistical indices of Classical test theory: item difficulty or p-value, discrimination (point biserial) analysis. Classical test theory as one of the measurement theories focuses on the true score of an examinee on a particular test. The theory predicts or explains the difficulty of questions, provides insights into the reliability of test scores, and help towards an assessment of how to improve the test by maintaining and developing a pool of good items from which future assessment tests could be drawn. Its essence is to assess the performance of individual test items on the basis of the overall quality of a test derived from the quality of its items. Classical test theory (CTT) is based on the premise that the observed score from a psychological testing is composed of an un-measurable true score and error. It is symbolically represented as: Observed score (X) = True score (T) + Measurement error (E), that is: ($X = T + E$)(Joshua, 2005; Adegoke, 2013; Bichi, Embong, Mamat, & Maiwada, 2015).The major issue in classical theory is on estimating the reliability of the observed scores of a test. Reliability is calculated through the individual's score on the test (observed score) and the amount of errors in the test itself (error), and together these give an indication of what the person's true score would have been without the errors in the test measurements. Errors in testing occur through systematic, administrative and other factors within the process. Classical theory holds that each individual has a true score which would be obtained if there were no errors in measurement. The implication of the classical test theory for test takers is that tests are fallible imprecise tools. The score achieved by an individual examine is rarely the individual's true score. The observed score is considered as the true score influenced by some degree of error. This error influences the observed scores to be higher or lower (Mango, 2009).

Difficulty index of each item in the test is expressed as the proportion of examinees that answered a particular item in the test correctly, expressed in p-value, as:

$$p = \frac{c}{n}$$

Where p = is the difficulty factor,
c = the number of examinees who gave correct responses to such item, and
n= total number of examinees.

Joint Committee on Standards for Educational and Psychological Testing of the AERA, APA, and NCME (2014) explained that p-values range from 0.0 to 1.0, with lower values corresponding to more difficult items and higher values corresponding to easier items. Joshua (2005), Thompson (2009) and Rana (2014) expressed that items difficulty values could be classified and interpreted as shown in Table 1

Table 1: Classification and Interpretation of Difficulty Indices

Percentage range	Difficulty index (p-value)	Quality	Interpretation
75 – 100	.75 – 1.0	Easy/Poor	Discard/Review
26 - 74	.26 - .74	Moderate	Retain
21 – 25	.21 - .25	Fair	Retain
20 and below	20 and below	Difficult	Discard

However, Zubairi, and Kassim (2006) and Adegoke (2013) opined that an item with a p-value of less than .30 is considered difficult; p –values of .31 to .70 expresses moderate difficulty, while p-value greater than .70 indicate easy item.

Item discrimination index measures how well an item differentiates between examinees with high and those with low abilities. There are several ways that discrimination could be computed: item discrimination power, mean item discrimination index and point-biserial correlation. The possible range of the discrimination index is considered as - 1.0 to 1.0. (Shakil, 2008; Denga, 2009). Point biserial is considered the most effective way to examine item discrimination, as every examinee who took the test is considered in the computation as against the upper 54% and lower 27 % which is the common case in discrimination computation. Point biserial is a product moment correlation that is capable of showing the predictive power an item has contributed to prediction by estimating the correlation between each item and the total test score of all the examinees (Triola 2006; Ghandi, Baloar, Alwi & Talib, 2013). The point –biserial correlation (r_{pbis}) is computed as follows:

$$r_{pbis} = \frac{Mpi - Mqi}{SD}\sqrt{piqi}$$

where:
Mpi = the mean total score of those answered item i correctly
Mqi = the mean total score of those answered item i incorrectly
SD = the standard deviation of all the exam
p = the proportion of group responding correctly to item i
q = the proportion of group responding incorrectly to itemi
The values of point biserial (r_{pbis}) for an item could be expressed as shown in Table 2.

Table 2. Classification and Interpretation of Point- Biserial indices

Point-biserial correlation coefficient (r_{pbis})	Quality
Below .19	Poor
.20 - .29	Marginal
.30 - .39	Good
.40 - .70	Very Good

Source: Penn, 2009; McGahee and Ball, 2009.

This study would use these classifications for difficulty and point biserial indices as a guide for analysis and decision.

Statement of the problem

The presence of problematic items in a test is a threat to reliability and validity, and makes the inference drawn about the examinees ability and proficiency illusive. Such if not adequately assessed, negates the properties of good test or measuring instrument which are: reliability, validity and usability of such results for any major decision making, locally or internationally. Further use of such item(s) in subsequent examinations without proper item analysis flaws effective and reliable assessment and a threat to quality educational enhancement.

Mock examination is conducted in Akwa Ibom state as a state final examination by the State Ministry of Education in preparing the senior secondary two (SS 2) students for national examinations conducted by West African Examination Council (WAEC) and National Examination Council (NECO), the examination bodies' saddled with the responsibilities of awarding the final certificates to the graduates of senior secondary school. The quality of each item in the test determines the extent of skills, proficiency and how well the students have acquired the expected mastery of the subject area to perform creditably well in external examinations or when called upon to demonstrate such skills. Because item analysis is one of the basic ways to ensure that items administered by the examination bodies are within the acceptable qualities. The researcher decided to examine the quality and characteristics of each of the 50 multiple –choice mathematics items that is administered to the senior secondary two (SS2) students by the Akwa Ibom State Ministry of Education in terms of difficulty levels, how well each item differentiates between high and low achievers, and extent of the relationship between the two indices.

Purpose of the study

The study was carried out to:

1. Examine the difficulty indices of each item: "moderate", "fair", "easy" and "difficult".
2. Examine the point-biserial indices of each item.
3. Examine the extent of relationship between various levels of difficulty indices with levels of point –biserial correlation coefficient indices.

Research questions

The following research questions were formulated to guide the study:

1. What are the various difficulty indices of each item of 2014 Akwa Ibom State Mock multiple –choice mathematics test?
2. What are the various point- biserial correlation coefficient indices of each item of 2014 Akwa Ibom State Mock multiple-choice mathematics test?
3. What is the relationship between the various levels of difficulty indices with levels of point –biserial correlation coefficient indices of each item in 2014 Akwa Ibom State Mock multiple-choice mathematics test?

Methodology

The study adopted an ex- post facto design. The choice of this design was considered appropriate in this study, since the researcher was only concerned with using the existing responses of the students after being dully administered and scored by examiners in the area. The researcher did not manipulate any of the variables in the study but only examined the quality and characteristics of each item. The population of the study consisted of 47, 600 senior secondary two (SS2) students' responses in 2014 State Mock Objective Mathematics in the public senior secondary schools in three educational zones of Uyo, Eket and Ikot Ekpene. 23,831 were males, while 23,768 were females. Sex of the participants is not considered in this study as the researcher is only considering item quality. The sample of the study comprised, 3,094 examinees' responses from the three educational zones randomly selected for the study. The 2014 State Mock multiple –choice Mathematics is a four optioned, A – D test. The data collection was done by the researcher through permission granted by the Director of the State Ministry of Education, Examination and Certification Unit, Akwa Ibom State. These responses were subjected to statistical analysis using IBM SPSS statistics version 20. The reliability of the test instrument was.71 Cronbach's Alpha based on standardized items.

Results

The results of the data analysis are presented in Tables 3, 4 and 5 according to the research questions

Research question 1: What are the various difficulty indices of each item of 2014 Akwa Ibom State Mock multiple –choice mathematics?

To answer this research question, the responses were subjected to IBM SPSS, 20. The output from the analysis on Table 3 indicates that 39 out of 50 items (5, 6, 7, 8, 9, 10, 11, 12, 13 14, 15,16, 17, 18,19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 41, 42, 43, 44, 45 and 46) are considered as moderately difficult items with p-value ranging from .71 - .74. Eleven items (1, 2, 3, 4, 38, 39, 40, 47, 48, 49 and 50) are classified as easy/poor items, with p- value ranges of .75 -.92. However, no item displayed fair and difficult indices.

Table 3: Item difficulty indices of 2014 Akwa Ibom State Mock Multiple –Choice Mathematics

Item No	P- value	Item No	P – value	Item No	P –value
1	.92	21	.72	41	.74
2	.83	22	.73	42	.74
3	.84	23	.74	43	.74
4	.77	24	.72	44	.71
5	.73	25	.72	45	.73
6	.74	26	.73	46	.73
7	.73	27	.72	47	.78
8	.74	28	.74	48	.81
9	.73	29	.73	49	.85
10	.71	30	.73	50	.88
11	.74	31	.73		
12	.74	32	.73		
13	.74	33	.73		
14	.72	34	.73		
15	.73	35	.71		
16	.72	36	.74		
17	.72	37	.74		
18	.71	38	.75		
19	.74	39	.75		
20	.72	40	75		

Easy item (.75 -1.0); Moderate item (.26 - .74); Fair item (.21 -.25) Difficult item (.20 and below)

Research question 2: What are the various point- biserial correlation coefficient indices of each item of 2014 Akwa Ibom State Mock multiple-choice mathematics?

To answer the research question, the output from SPSS was examined. The result on Table 4 shows that 32 items (1, 3, 9,10, 11, 13, 14, 15, 16, 17,19, 21, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 34, 35, 37, 39, 40, 41, 42, 43, 44 and 47) have point-biserial correlation coefficients below .19, indicating that such items display poor discrimination between the high and low achievers. Four items (2, 8, 20 and 50) have point-biserial of .19. 14 items (4, 5, 6, 7, 12, 20, 33, 36, 38, 45, 46, 48, and 49) have marginal point- biserial correlation coefficient of .20 to .22. The results indicate that only few items could be said to possess marginal discrimination power between high and low achievers.

Table 4: Item Point- Biserial Correlation Coefficient Indices of 2014 Akwa Ibom State Mock Multiple –Choice Mathematics

Item	Point- Biseerial (r_{pbis})	Item	Point- Biseerial (r_{pbis})	Item	Point- Biseerial (r_{pbis})
1	.14	21	.17	41	.16
2	.19	22	.16	42	.17
3	.13	23	.17	43	.13
4	.20	24	.18	44	.16
5	.21	25	.20	45	.20
6	.22	26	.17	46	.20
7	.20	27	.17	47	.18
8	.19	28	.15	48	.20
9	.18	29	.16	49	.20
10	.18	30	.18	50	.19
11	.18	31	.18		
12	.20	32	.17		
13	.15	33	.20		
14	.17	34	.17		
15	.14	35	.18		
16	.18	36	.20		
17	.21	37	.17		
18	.18	38	.20		
19	.16	39	.18		
20	.19	40	.16		

.19 and below (poor discrimination), .20 – 22 (marginal discrimination).

Research question 3: What is the relationship between various levels of difficulty indices with level of point –biserial correlation coefficient indices of each item of 2014 Akwa Ibom State Mock multiple-choice mathematics?

This research question was answered using output from SPSS 20.The result reveals that good items with p-value range of .71 to .74 displayed poor point biserial correlation coefficient of .19 and below, except items nos.5, 6, 7, 12, 17, 25, 33, 36, 45 and 46 displayed marginal point- biserial correlation coefficient of .20 to .22. The easy items of 1, 2, 3, 39, 40, 47 and 50 with p-value range from .75 to .92, have poor point- biseerial correlation coefficients of .19 and below, except items nos: 4, 38, 48 and 49, displayed marginal point –biserial correlation coefficient value of .20. This result indicates that even those items within moderately p-value range of .71 - .74, lack the potential to differentiate the high from the low scorers. Also, some of the easy items are not even good discriminators, except very few. Thus, the relationship is not linear.

Table 5: Relationship between Various Levels of Difficulty Indices with Levels of Point –Biserial Correlation Coefficient Indices of each Item of 2014 Akwa Ibom State Mock Multiple-Choice Mathematics.

Item No	P-value	r_{pbis}	Item No	P- value	r_{pbis}	Item No	P- value	r_{pbis}
1	.92	.14	21	.72	.17	41	.74	.16
2	.83	.19	22	.73	.16	42	.74	.17
3	.84	.13	23	.74	.17	43	.74	.13
4	.77	.20	24	.72	.18	44	.71	.16
5	.73	.21	25	.72	.20	45	.73	.20
6	.74	.22	26	.73	.17	46	.73	.20
7	.73	.20	27	.72	.17	47	.78	.18
8	.74	.19	28	.74	.15	48	.81	.20 .
9	.73	.18	29	.73	.16	49	.85	.20
10	.71	.18	30	.73	.18	50	.88	.19
11	.74	.18	31	.73	.18			
12	.74	.20	32	.73	.17			
13	.74	.15	33	.73	.20	.		
14	.72	.17	34	.73	.17			
15	.73	.14	35	.71	.18			
16	.72	.18	36	.74	.20			
17	.72	.21	37	.74	.17		.	
18	.71	.18	38	.75	.20			
19	.74	.16	39	.75	.18			
20	.72	.19	40	.75	.16			

1. **Easy item (.75 -1.0); Moderate item (.26 - .74); Fair item (.21 -.25) Difficult item (.20 and below). 2. .19 and below (poor discrimination) .20–22(marginal discrimination).**

Discussion of Findings

The results on the difficulty indices of the State Mock examination showed that 38 items were considered as moderate items according to the p-values that ranged from .71 to .74. Eleven items were considered as having easy/poor difficulty indices ranging from .75 to .92. However, no item was considered as hard or difficult. The findings are in line with what Ghandi, et al (2013) and Rana (2014) found out in different studies carried out at Haryana and Putra, Malaysia on assessing difficulty indices. The researchers found that good items displayed p-value range from .40 to .71, and .30 to .76 respectively. However, items in this study lack spread of difficulty indices, as items were clustered without a spread throughout the test. For instance, items with moderate difficulty indices spanned from 5 to 37 and 41 to 46; with easy items clustering between 1 to 4, 38 to 40, and then 47 to 50. This result contradicts the finding by Ghandi, et al (2013) who discovered spread of items in the study with good, easy and difficulty questions adequately spread across the test. Furthermore, the implication of the results on the difficulty indices according to Thompson (2009) and Rana (2014), on difficulty values showed that the 2014 Mock multiple mathematics paper does not have a spread of items to cover the various difficulty levels, and such arrangement of items is liable to promote guessing. With only two difficulty indices of moderate and easy items as indicated by the p-values, an effective review of this instrument is imperative.

On point biserial correlation coefficient indices, the result revealed a marginal discrimination power with 14 items having point-biserial correlation coefficient that ranged from .20 to .22. 4 items had point-biserial correlation coefficient of .19, while 32 items had point –biserial correlation coefficient of below .19, indicating poor point biserial correlation coefficient. This indicated that 36 items failed to discriminate effectively between the high and the low scorers. However, this study agrees with finding by Sim and Rasiah (2006) in a study that revealed a minimum discrimination power of items. But Verma (2008) in a study stressed that a low point-biserial correlation coefficient results is an indication that students who got the item incorrect also scored high on the test, while those who got the item correct scored low on the overall test. Furthermore, Popham (2008) emphasized that items with marginal point biserial correlation coefficient may need adjustment, while items with .19 and below should be rejected or removed from the test. Thus, items with low point biserial values need further examination or an outright removal from scoring and future testing. According to Karelia, Pillai, Vegada (2013), some of the common causes of poor discrimination in test items may be, ambiguous wording and wrong keys.

The results on the relationship between the difficulty and discrimination (point-biserial) indices showed that some of the good items failed to meet the acceptable standard of discrimination for this study, which is .20, except 10 items earlier listed. Also, some of the easy items as well had point –biserial correlation coefficient of.19 and below. Only few items in the test had marginal point- biserial correlation coefficient of .20. The findings of this study agrees with the study carried out by Sim and Rasiah (2006) on the relationship between difficulty and discrimination indices in true/false type multiple choice question of a para-clinical multidisciplinary paper, which revealed that relationship between the difficulty and discrimination indices of the items were not linear but dome-shaped, as marginal discrimination occurred with moderately easy items. Furthermore, Karelia, Pillai, Vegada (2013), found in a study that Pearson correlation between difficulty and discrimination indices showed that discrimination index correlated poorly with difficulty index (r =.11) with insignificant p-value at .05 alpha level, though the researchers used the 54% upper and the 27% lower discrimination method. In the same vein, Mitra, Nagaraja, Ponnudurai and Judson (2009) in a study found that discrimination correlated poorly with difficulty index (r.3) which signified that with an increasing difficulty index values, discrimination index witnessed decreasing values. The result indicated that low performance students were disposed to getting more correct answer.

Conclusion

The study revealed that some items were classified as having moderately difficulty index with acceptable p-values. Few items were considered as easy or poor with very high p-values. The point-biserial correlation coefficients of some items were poor, while others items had marginal point-biserial correlation coefficient. It is therefore concluded that analysis of difficulty and point- biserial indices of items as a means of ensuring quality assurance of test items administered in education is imperative for reliable and valid test data.

Recommendations

Based on the findings, the following suggestions were made that:

1. The test developers and administrators of the State Mock examination should carry out pilot testing of items that make the test to ensure the quality of items that meet the required standard of difficulty and discrimination indices.
2. The 2014 multiple–choice mock mathematics should be reviewed to diversify the difficulty levels and spread the items adequately.
3. The standard used in this study for the difficulty and discrimination indices should be used by the Examination and Certification Unit of the Ministry of Education for

carrying out item analysis that would provide a guide for developing and administering items that are acceptable and good for the purpose.

4. Frequent items analysis of the State Mock papers should be carried out to ensure dependable and reliable item bank for future item use.

References

Adegoke, B.A. (2013).Comparison of item statistics of Physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice*, *4*(22), 87 -96.

Bichi, A.A., Embong, R., Mamat, M., & Maiwada, D.A. (2015). Comparison of classical test theory and item response theory : A review of empirical studies. *Australian Journal of Basic and Applied Science*, *9*(7), 549-556.

Denga, I. (2009).*Educational measurement, continuous assessment and psychological testing*.Rapid Educational Publishers Ltd. Calabar, Cross River State, Nigeria.

Ghandi,, I. N., Baloar, A. K., Alwi, N.H., &Talib, O. (2013). Measuring critical thinking skills of undergraduate students in University Putra Malaysia. *International Journal of Asian Social Science, 3* (6), 1458- 1466.

Joint Committee on Standard for Education and Psychological Testing of AERA, APA, NCME (2014).Retrieved from: http/www.teststandards.org/files/standards.

Joshua, M.T. (2005). *Fundamentals of test and measurement in education.*University of Calabar Press, Calabar, Nigeria.

Karelia, B.N., Pillai, A., & Vegada, B. N. (2013). The levels of difficulty and discrimination indices and relationship between them in four- response type multiple choice questions of pharmacology summative test of Year 11 M.B.B.S students.*IcJSME*,*7*(2), 41-46.

Krishnan, V.(2013). The early child development instrument (EDI) : An item analysis using classical test theory (CTT) on Alberta's data. Retrieved from: http://www.cup.ualberta.ca/wp-content/uploads/2013/04/ItemAnalysisCTTCUPWebsite_.pdf

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*,1(1), 1-11. Retrieved from: http://hbanaszak.mjr.uw.edu.pl/TempTxt/Magno+2009.

McGahee, T. W., & Ball, J. (2009). How to read and really use an item analysis. *Nurse Educator, 34*, 166 – 177.Retrieved from: http/www.creighton.edu/sites/www1

Mitra, N.K., Nagaraja, H.S., Ponnudurai, G., & Judson, J.P. (2009). The levels of difficulty and discrimination indices in type A multiple choice questions of pre- clinical Semester 1 multidisciplinary summative tests. *IcJSME*,*3*(1)*, 2 -7.*

Penn, B. K. (2009). Test item development and analysis. Presented at Creighton University School of Nursing Faculty Retreat, Omaha, NE. Retrieve from: http/www.creighton.edu/sites/www1.

Popham, J. W. (2008). *Classroom assessment: What teachers need to know*. Boston: Pearson Education, Inc.

Rana, S.S (2014). Test item analysis and relationship between difficulty level and discrimination index of test items in an achievement test in Biology. *Indian Journal of Research,3* (6), 56 – 58.

Shakil, M. (2008).Assessing student performance using test item analysis and its relevance to the state exit final exams of MAT0024 classes: An action research project.Retrieved from:http/www.mdc.edu/main/imafes/Usi.

Sim, S. S., &Rasiah, R I. (2006). Relationship between item difficulty and discrimination indices in true/false type multiple choice questions of a para – clinical multidisciplinarypaper. *Ann Acad Med*., 3 , 67 -71.

Thompson, N.A. (2009). Classical item and test analysis with CITAS. *Assessment system Corporation*, 1 – 8. Retrieved from :http/assess.com/docs/Thomps.

Triola, M. F. (2006). *Elementary statistics*. Pearson Addison – Wesley, New York.

Verma.S. (2008). Preliminary item statistics using point – biserial correlation and p-values. Retrieved from : http/ www.eddata.com/resources,publications/EDS Point Biserial (pdf).

Zubairi, A.M. & Kassim, N. L. A. (2006). Classical and Rasch analysis of dichotomously scored reading comprehension test items. *Malaysia Journal of ELT Research*, 2, 1-20.