

## **Empirical Demonstration of Techniques for Computing the Discrimination Power of a Dichotomous Item Response Test**

B. O. Ovwigho

*Department of Agricultural Economics and Extension,  
Delta State University, Asaba Campus, Asaba, Nigeria*

---

**Abstract:** *The study was designed to examine the various statistical techniques for computing the Discrimination Power of a dichotomous item response test. The simple Discrimination Index, the Cramer's phi coefficient, phi coefficient, Point-Biserial correlation, and t-test for independent sample were used to test the difference in mean of the criterion scores between those who got the item correct and wrong. Item 1 of the Senior Secondary Agricultural Science test was used as the computational example. In the analysis, the simple Discrimination Index and 3 formulae of the Point-Biserial correlation gave the same results ( $r = 0.63$ ). The Cramer's phi coefficient was 0.49 and phi coefficient (0.65). There was a significant difference between the mean criterion scores of students who got the item right and those who got the item wrong ( $t = 5.05$   $p < 0.05$ ). The phi coefficient, Point-Biserial correlation and t-test for independent sample were the best statistics for estimating the Discrimination Power and coefficient of the test item. The simple Discrimination Index did not take into cognisance the total performance of candidates in the test hence it was rejected. The phi coefficient is recommended where the sample is large and necessitates division into upper and lower extreme groups.*

**Keywords:** *Item analysis, dichotomous item, discriminating power, discrimination index, discrimination coefficient*

---

### **I. Introduction**

#### **BACKGROUND INFORMATION**

Two types of tests: criterion referenced tests and norm referenced test are encountered in testing programmes. The criterion referenced test is used to determine whether a candidate possessed the quality being measured while the norm referenced test measures the relative performance of the candidate in a group. Norm referenced tests are used more frequently by teachers. The teacher can make or mar the educational career of a student by the quality of test he administers to examinees. To guide against poor testing of students, various methods and techniques have been developed for improving the quality of tests. The various techniques for improving the quality of test items in the social and behavioural sciences fall under the umbrella name often referred to as item analysis. Generally item analysis encapsulates statistical techniques for improving the quality of test items.

According to Instructional Assessment Resources (IAR 2011), Matlock-Hetzel (1997), Ali *et. al.* (1988) and Gronlund (1976) item analysis involved many statistics that could provide useful information for improving the quality and accuracy of multiple or true-false items (questions). The item analysis procedures include estimate of item difficulty, item discrimination, reliability coefficient, and distractor evaluation. Item difficulty, also known as  $p$ -value, was calculated by dividing the number of students who got an item correct by the total number of students who answered it. However this paper is concerned only with the construction of discriminating power of dichotomous item response test.

Ovwigho (2011), IAR (2007) and Matlock-Hetzel (1997) stated that item discrimination was the relationship between how well students did on the item and their total examination scores. They noted that if the test and a single measures the same thing one would expect students who did well on the test to answer that item correctly and those who did poorly to answer the item incorrectly. A good test item discriminates between those who did well and those who did poorly. The higher the value the more discriminating is the item. A highly discriminating item indicated that students with high scores in the examination got the item correct whereas students who had low scores got the item wrong. Items with discrimination values near or less than zero should be removed from the examination because it showed that students who did poorly in the examination did better on the item than students who performed well in the examination. The value ranges between -1.00 to 1.00. Ebel (1972) stated various values of discrimination Indices and how to evaluate them (Table 1).

**Table 1: Evaluation of discrimination indexes after 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, R. L. (1972). *Essentials of Educational Measurement*

Matlock-Hetzel (1997) noted that two types of indices could be computed to determine the discriminating power of a test item- the Discrimination Index and Discrimination Coefficient.

**Discrimination Index** This is also referred to as the simple discrimination index. Matlock-Hetzel (1997), Gronlund (1976) and Ali *et. al* (1988) stated that the extreme group technique could be applied to compute the Discrimination Index. The time tested procedures were:

- Score each of the test answer papers and rank order the scores or arrange from high to low
- Separate the examinees into two groups made up of an upper 27% of the total group who received highest scores on the test and lower group of 27% of the total group who received lowest scores on the test> This agreed with Wiersma and Jurs (1990) that the upper and lower 27% maximised differences in normal distribution and it provided enough cases for analysis
- Count the number of times each possible response to an item was chosen on the papers of the upper group. Do the same thing for the papers of the lower group
- Record these response counts for each item
- Subtract the lower group count of correct responses from the upper group count of correct responses. Divide the difference by the maximum possible difference (the number of papers in the upper or lower group). The quotient is referred to as Discrimination Index. According to them the Discrimination Index could be expressed mathematically as:

$$DI = \frac{RU - RL}{1/2T} \dots\dots\dots (1)$$

where

DI = Discrimination Index;

RU = Number of persons in the upper group who got the item right;

RL = Number of persons in the lower group who got the item right; and

1/2T = Half the total number of persons in both upper and lower groups.

**Discrimination Coefficient** Many statistical techniques are available for calculating the discrimination coefficient. The Cramer’s phi coefficient, Point-Biserial Correlation and Pearson Product Moment Correlation were computed for the same sample and item

The Cramer’s phi coefficient is an extension of the Chi square test the strength of association between two variables. Joe (1992) stated that the Cramer’s phi coefficient was used to test the association between two nominal variables if the Chi square ( $X^2$ ) was known. The formula was expressed as:

$$\phi = \sqrt{\frac{x^2}{N(L-1)}} \dots\dots\dots 2$$

Where:

$X^2$  = Chi square value;

N = Number of Frequencies; and

L = Number of rows or columns, whichever is smaller

Source: Joe, A. I. (1992). *Fundamental Statistics for Education and the Behavioural*

*Sciences*

The Point- Biserial correlation is advantageous over the Discrimination Index and the phi coefficient because it takes into cognisance the score of every candidate in the computation of the Discrimination Coefficient. Matlock-Hetzel (1997) stated that the Point-Biserial correlation was used to find out if the right people got the item right and the predictive power of the item. Henrysson (1971) stated that that Point-Biserial correlation was concerned with item-criterion relationship and predictive validity of the total test. Ovwigho (2009) modified the Point – Biserial correlation from Henrysson (1971) to calculate the indexes of dichotomous socio-economic status indicators. The modified formula was as follows:

$$rpbis = \frac{MP - MN}{ST} \cdot \sqrt{p(1 - p)} \dots\dots\dots 3$$

where

rpbis = Symbol for Point – Biserial correlation;

$MP$  = Mean criterion score for heads of farm families who possessed the item;  
 $MN$  = Mean criterion score for heads of farm families who did not possess the item;  
 $ST$  = Standard deviation of the criterion scores; and  
 $P$  = Proportion of heads of farm families who possessed the item.

Source: Ovwigho, B.O. (2009). Validation of Socio-economic Status Indicators

However, Adesoye (2004) expressed the same formula as shown in equation 4

$$rpbis = \frac{MP-MQ}{ST} \cdot \sqrt{PQ} \dots\dots\dots 4$$

Where:

$MP$  = Mean of the continuous data for group P;  
 $MQ$  = Mean of the continuous data for group Q;  
 $St$  = Standard deviation of the entire population;  
 $P$  = Proportion of group P; and  
 $Q$  = Proportion of group Q

Source: Adesoye, P. O. (2004). *Practical Guide to Statistical Analysis for Scientists*.

IAR (2011) stated that item discrimination was also referred to as the Point-Biserial correlation (PBS). The formula they expressed was slightly different from equation 3 and 4 (see equation 5).

$$r = \frac{(XC-XT)}{SD\ Total} \cdot \sqrt{pq} \dots\dots\dots 5$$

where

$XC$  = the mean total score for persons who got the item correct;  
 $XT$  = mean total score for all candidates;  
 $P$  = difficulty value for the item;  
 $Q$  = (1 - P); and  
 $SD_T$  = Standard Deviation of total examination scores

Source: IAR (2011). Item Analysis

Wikipedia (2012) stated that the Point-biserial correlation coefficient (rpb) was mathematically equivalent to the Pearson (product moment) correlation. It is applicable when we have one continuously measured variable  $X$  and a dichotomous variable  $Y$ . This could be shown by assigning two distinct numerical values to the dichotomous variables. The Point-biserial correlation coefficient could be calculated as follows:

$$rpb = \frac{M1-Mo}{Sn} \cdot \sqrt{\frac{n1n0}{n^2}} \dots\dots\dots 6$$

Where:

$M1$  = Mean value of the continuous variable group 1;  
 $Mo$  = Mean value of the continuous variable group 2;  
 $n1$  = Number of data points in group 1;  
 $no$  = Number of data points in group 2;  
 $n^2$  = Total sample size squared; and  
 $Sn$  = standard deviation for every member of the population

Source : Wikipedia (2013). Point-biserial Correlation Coefficient

The Pearson Product Moment Correlation Coefficient could be used to calculate the discrimination index of dichotomous item response. However the Pearson r coefficient may be spurious if the two continuous samples are not equal. Thus the t-test for independent sample could be used to test the difference in mean. This could be expressed mathematically as:

$$t = \frac{M1-M2}{\sigma\ M1-M2} \dots\dots\dots 7$$

Where:

$M1$  – Mean for sample 1; and  
 $M2$  – Mean for sample 2;

$\sigma\ M1 - M2$  = Standard error of the difference between the means

Where:

$$\sigma\ M1 - M2 = \sqrt{\frac{SS1+SS2}{N1+N2-2} \cdot \frac{N1+N2}{N1N2}} \dots\dots\dots 8$$

Where:

$SS1$  and  $SS2$  = Sum of squares for sample 1 and 2 respectively

Where:

$$SS = \sum x^2 + \frac{(\sum x)^2}{N} \dots\dots\dots 9$$

Source: Joe, A. I. (1992). *Fundamental Statistics for Education and the Behavioural Sciences*

The foregoing exemplified the variegated formulae used for computing discrimination indexes and coefficients. Thus the study was designed to demonstrate some of these statistical techniques for calculating the discrimination index and coefficients of dichotomous item response test. The specific objectives were to:

- i calculate the simple Discrimination Index of the dichotomous item;
- ii use the Cramer's phi coefficient and conventional phi coefficient in estimating the discrimination coefficient of a dichotomous item;
- iii use and compare different formulae of Point Biserial Correlation to calculate the discrimination coefficient of a dichotomous item; and
- iv apply the t-test for independent sample to the difference in mean criterion score between those who got the item correct and wrong

## II. Computational Example and Discussion

The sample data was made up of a test containing 20 dichotomous and quantitatively measured items administered to 40 Senior Secondary School 2 Agricultural Science students. After scoring the entire test, the discrimination index and coefficients of item 1 was calculated by various statistical techniques. Item 1 was a dichotomous item which was captioned as:

Cattle is a ruminant animal. True or False?

The criterion scores which represented the overall scores of the students in the test and the number of correct and wrong responses to the item were presented in Table 1. The criterion scores are arranged in a descending order. The individuals who answered the item correctly or wrongly and their respective criterion scores were shown in the Table columns.

**Discrimination Index** 27% = 11; Upper 27% who got the item correct = 10; lower 27% who got the item correct = 3

$$\frac{1}{2} \text{ total} = 11$$

From equation 1 the Simple Discrimination Index could be calculated as:

$$DI = \frac{10-3}{11}$$
$$DI = 0.63$$

The disadvantage with the discrimination index is that it does not take every subject into consideration in the analysis. It could be useful where the sample size is large as a quick check of the discriminating power of the test. In this example the DI had the same value with the Point-Biserial correlation

**Cramer's Phi Coefficient** In applying the Cramer's phi coefficient to the data, the Chi square was first calculated (Table 2).

In parenthesis = Expected Frequency

The Cramer's phi coefficient could be calculated by fitting the data into equation 2 as follows:

$$\phi = \sqrt{\frac{5.20}{22(2-1)}}$$
$$\phi = 0.49$$

**Point-Biserial Correlation:** Formulae 3, 4, 5 and 6 were applied to find the Point-biserial correlation of the data. Formulae 3, 4 and 6 gave the same result.

$$P = \frac{30}{40} = 0.75 \quad MC = 58.23 \quad MW = 38.50 \quad SD = 13.60$$

$$r_{pbis} = \frac{58.23-38.50}{13.60} \cdot \sqrt{0.75(1-0.75)}$$
$$r_{pbis} = 0.63$$

Matlock-Hetzel (1997) noted that the point biserial ( $r_{pbis}$ ) correlation was used to find out if the right people got the items right, and how much predictive power the item has and how it would contribute to predictions

The formula for Point-Biserial correlation offered by IAR(2011) shown in equation 5 was tested.

$$r = \frac{(58.23-53.30)}{13.60} \cdot \sqrt{0.75(1-0.75)}$$
$$r = 0.157$$

The coefficient was spurious hence it should not be accepted for calculating Point-Biserial correlation. The mean of the wrong responses should be used instead of the mean of all candidates as contained in the formula. The Difficulty value (75.00%) of the whole test is different from the upper and lower 27% difficulty (59.09%). The difference stems from the fact that not all the candidates were used in the latter case. It is better to get closer or use the entire population in statistical analysis in order to get a better picture of the parameters of interest.

**The Pearson r Phi Coefficient and t-test for Independent Sample:** The Pearson r of the data gave a spurious result because the two samples were unequal. The alternative was to apply the Phi Coefficient procedure

described by Wikipedia (2013) and Calkins (2005). They stated that the Phi Coefficient was a simplified way of calculating the Pearson Product Moment Correlation. They gave the formula as:

$$\phi = \frac{ad-bc}{\sqrt{efgh}} \dots\dots\dots 10$$

Thus the Phi Coefficient of the sample data in Table 2 could be calculated as follows:

$$\phi = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \\ \phi = \frac{77}{118.98} \\ = 0.65$$

The phi coefficient (0.65) is closer to rpbis (0.63) and DI (0.63) hence it should be accepted as a valid measure for calculating the Discrimination Coefficient of test item

The t-test for unpaired samples was applied to compare the mean criterion scores of the candidates who got the item right and those who got it wrong (Table 3).

MC = 58.23 MW = 38.50 SS1 =2507.37 SS2 = 1968.50

By applying equation 7 and 9, the t-value could be calculated as follows:

$$t = \frac{58.23-38.50}{\sqrt{\frac{2507.37+1968.50}{38} \cdot \frac{40}{300}}} \\ t = 5.05; p < 0.05$$

There was a significant difference ( $p < 0.05$ ) between the criterion scores of students who got the item correct and those who got it wrong. This meant that the right people got the item correctly. Thus the item could be accepted as a valid discriminator between high or clever and low or dull students. In addition the item could be used to predict the overall performance of a student in the test.

### III. Conclusion And Recommendations

The Discriminating Power of test items could be measured by the Discrimination Index and Discrimination Coefficient. Social and behavioural scientists should go beyond the simple Discrimination Index in estimating the Discriminating Power of test items. The simple Discrimination Index could be used as a quick check of the Discrimination Coefficient of a test item. The Cramer's Phi Coefficient, Pearson r and formula by Instructional Assessment Resource IAR (2007) did not give good estimate of the Discrimination Coefficients of the test item. The Point-Biserial correlation, phi coefficient and t-test gave the best estimates of the Discrimination coefficient of the test item. The phi coefficient is recommended where the sample is large and divided into upper and lower extremes.

### References

- [1]. Adesoye, P. O. (2004). *Practical Guide to Statistical Analysis for Scientists*. Ibadan : Debo Publishers
- [2]. Ali, A., Ezeadi, S. and Ogbazi, J. N. (1988). *Introduction to Educational Measurement*. Awka(Nigeria): Meks Unique Publishers
- [3]. Calkins, K. G. (2005). *Applied Statistics Lesson 13: More Correlation Coefficients*. Michigan United States: Andrews University, Berrien Springs
- [4]. Ebel, R. (1972). *Essentials of Educational Measurement*. London: Prentice Hall
- [5]. Gronlund NE 1976, *Measurement and evaluation in teaching*, Macmillan Publishing Co., Inc. New York
- [6]. Henrysson, S.(1976). *Gathering, Analyzing and Using Data on Test Items*. In: Thorndike, R. D. (ed). *Educational Measurement*. Washington DC: American Council on Education
- [7]. Instructional Assessment Resources, IAR (2011). *Item Analysis*. Texas: University of Texas at Austin
- [8]. Joe, A. I. (1992). *Fundamental Statistics for Education and the Behavioural Sciences*. Ibadan: Kraft Books Limited
- [9]. 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
- [10]. Ovwigbo, B. O. (2009). Validation of Socioeconomic Status Indicators for Heads of Farm Families in Delta north Agricultural Zone, Delta State, Nigeria. *Journal of Agricultural Extension and Rural Development*, 1(2): 50-57
- [11]. Ovwigbo, B. O. (2011). Construction of Socio-economic Status Indexes for Heads of Rural Farm Families in the Central Agricultural Zone of Delta State, Nigeria. *Extension Farming Systems Journal*, 7(1): 21-28
- [12]. Wiersma, W. and Jurs, S. G. (1990). *Educational Measurement and Testing*. Boston : MA Allyn and Bacon
- [13]. Wikipedia (2013). Point -biserial Correlation Coefficient. Retrieved from [en.wikipedia.org/wiki/Point-biserial\\_correlation\\_coefficient](http://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient) on 10<sup>th</sup> January, 2013
- [14]. Wikipedia (2013). Phi Coefficient. Retrieved from [en.wikipedia.org/wiki/Phi\\_-coefficient](http://en.wikipedia.org/wiki/Phi_-coefficient) on 10<sup>th</sup> January, 2013

Table 1: Criterion scores, proportion of right and wrong answers

S/N	Criterion Scores	Number Correct	Number Wrong
1	76	1	
2	75	1	
3	74	1	
4	72	1	
5	71	1	
6	70	1	
7	68	1	
8	66		1
9	65	1	
10	63	1	
11	60	2	
12	59	1	
13	58	1	
14	57	1	
15	56	1	
16	55	1	
17	54	3	
18	53	1	
19	52	2	
20	50	6	1
21	49	1	
22	48		1
23	45		1
24	44		1
25	40	1	
26	35		1
27	30		1
28	28		1
29	20		1
30	19		1
Total	2132	30	10

Table 2: Contingency Table of upper and lower bound of responses

Bound/	Upper	Lower	Total
Response categories			
Correct	A 10(6.5)	B 3(6.5)	13
Wrong	C 1(4.5)	D 8(4.5)	9
Total	11	11	22

$x^2 = 5.20; p < 0.05$

Table 3: Distribution of criterion scores according to right and wrong answer to the item

Criterion Scores of Candidates who got item correct	Criterion Scores of Candidates who got the item wrong
76	66
75	50
74	48
72	45
71	44
70	35
68	30
65	28
63	20
60	19
60	
59	
58	
57	
56	
55	
54	
54	
54	
53	
52	
52	
50	
50	
50	
50	
50	
50	
49	
40	