IJPES
www.ijpes.com
International Journal of Psychology and Educational Studies
ISSN: 2148-9378

Assessment of Item and Test parameters: Cosine Similarity Approach

Satyendra Nath CHAKRABARTTY[1]

[1]*Indian Statistical Institute, Pradesh, India* *0000-0002-7687-5044*

ARTICLE INFO

Article History
Received 29.11.2020
Received in revised form 28.03.2021
Accepted 09.06.2021
Avaible Online: 25.07.2021
Article Type: Research Article

ABSTRACT

The paper proposes new measures of difficulty and discriminating values of binary items and test consisting of such items and find their relationships including estimation of test error variance and thereby the test reliability, as per definition using cosine similarities. The measures use entire data. Difficulty value of test and item is defined as function of cosine of the angle between the observed score vector and the maximum possible score vector. Discriminating value of test and an item are taken as coefficient of variation (CV) of test score and item score respectively. Each ranges between 0 and 1 like difficulty value of test and an item. With increase in number of correct answer to an item, item difficulty curve increases and item discriminating curve decreases. The point of intersection of the two curves can be used for item deletion along with other criteria. Cronbach alpha was expressed and computed in terms of discriminating value of test and item. Relationship derived between test discriminating value and test reliability as per theoretical definition. Empirical verifications of proposed measures were undertaken. Future studies suggested.re to enter text.

Keywords:
Difficulty values; discriminating values; cosine similarity; coefficient of variation; reliability.

1. Introduction

Tests consisting of binary items are traditionally scored as 1 for right answer and 0 for rest. Such scoring are frequently used for assessment in different educational levels. Item analysis aims at assessing the quality of the constituent items and test as a whole by revising or discarding ineffective items. Two popular measures are item difficulty value and item discriminating value. Difficulty value of an item is defined as the proportion of correct responses to the item. Higher difficulty value implies the item was easy and higher discriminating value implies that the item was more able to discriminate between students of higher and lower abilities. Item discriminating value refers to the ability of an item to distinguish between examines with high ability level from those with low ability level (Ferrando, 2012). Discriminating value of a binary item is traditionally computed as the upper-lower index using top 27% and bottom 27% of data and rejecting 46% of the data and hence may not be desirable. Moreover, relationship between item difficulty values ($Diff_i$), based on the entire data and item discriminating values ($Disc_i$) based on 54% of the data is not straight forward and have resulted in contrasting results. For example, Rao, et al. (2016) found positive correlation (0.563) between $Diff_i$ and $Disc_i$. Sim and Rasiah (2006) found that $Diff_i$ and $Disc_i$ are correlated positively at the "easy end" (where percentage difficulty values ranged between 80% and 100%), but negatively at the "difficult end" (where percentage difficulty values were between 0% and 20%) and dome-shaped curve when all items are considered. The authors suggested for evaluation of effectiveness of MCQ items. Chauhan, et al. (2013) proposed further study to investigate correlation between difficult index and discriminative index. Researchers differed marginally on the cutting points of classification of items under "poor discrimination power", "excellent discrimination", "good discrimination", etc. Lack of relationship between $Diff_i$ and $Disc_i$ and their relationships with test

[1] Corresponding author: Indian Statistical Institute, Pradesh, India
e-mail: chakrabarttysatyendra3139@gmail.com
Citation: Chakrabartty, S. N. (2021). Assessment of item and test parameters: Cosine similarity approach. *International Journal of Psychology and Educational Studies*, *8*(3), 28-38.
https://dx.doi.org/10.52380/ijpes.2021.8.3.190

parameters could not reflect impact of deletion of one or more items on test reliability (r_{tt}) or error variance ($S_E{}^2$) or discriminating value of the test ($Disc_T$) or difficulty value of the test ($Diff_T$).

Reliability coefficient does not serve the purpose of quantifying the degree of discrimination offered by an instrument (Hankins, 2007). Inclusion of an item with negative or zero discrimination may result in measurement disturbance regarding the test. Thus, discriminating value is directly related to the quality of the score as a measure of the trait (McDonald, 1999). Item discriminating values are usually lower for non-homogeneous tests. Range of the discrimination index is between - 1.0 to 1.0. (Shakil, 2008; Denga, 2009).

Moreover, to assess quality of test as a whole, it is needed to consider test parameters like difficulty value and discriminating value of the test and find their relationships with other parameters like test reliability, validity. Discriminating value of a test is a test characteristic which is different from reliability and validity. One of the major objectives of a test is to find how the test can discriminate good performers from others or to see the extent to which an item or the entire test can discriminate the sample. The objective can be achieved if we find discriminating value of an item and discriminating value of a test.

Approaches without ignoring significant percentage of data include item-total correlation, bi-serial correlation (r_{bs}) , point bi-serial correlation (r_{pbs}), Spearman's correlation, etc. between item score and test scores (with or without that item) (Tzuriel and Samuels, 2000). While r_{bs} describes the relationship between an item score and scores on the total test for all examinees (Ebel and Frisbie, 1991), r_{pbs} reflects the predictive validity of the test (Henrysson, 1971). Moreover, r_{bs} tends to favor items of average difficulty. Researchers tend to differ on cutting point value of item-total correlation, below which items may be deleted. For example, Kehoe (1995) suggested restructuring of the items which have item-total correlation less than 0.15 since such items do not measure the same ability as does the test. But, Popham (2008) suggested rejecting the items for which $r_{pbs} \leq 0.19$

Need is to have reliable method of computing difficulty value and discriminating value of a test and items and find their relationships with test reliability under classical test theory (CTT). The complex model of Item Response Theory (IRT) was not considered primarily for its requirement of large sample size and strict assumptions including a curvilinear relationship between item score and construct score against a simple linear relationship between them by CTT.

The paper gives methods of obtaining difficulty and discriminating value of items and also tests using angular similarities and their relationships including estimation of test error variance and thereby the test reliability, as per definition (ratio of true score variance and observed score variance), via a single administration without sacrificing any portion of data and making no assumption of continuous nature or linearity or normality for the observed variables or the underlying variable being measured. Thus, the approach is an improvement over observation made by Rudner and Schafes, (2002) who mentioned that it is impossible to calculate a reliability co-efficiency that conforms to the theoretical definition since true scores of individuals taking the test are not known.

Rest of the paper is organized as follows. In the following Section, the proposed methodology of obtaining difficulty, discriminating values of binary items and test consisting of such items under CTT is elaborated along with derivation of relationship between difficulty values and discriminating values of an item and other parameters like item reliability and test reliability. Details of the empirical verification for the proposed methods are discussed in Section 3. The paper is rounded up in Section 4 by recalling the salient outcomes of the work.

2. Method

Consider a test consisting of *m*-binary items (1 for correct answer and 0 otherwise) has been administered to *n*-respondents, where $n > m$. Let $\mathbf{X}=(X_1, X_2, \ldots, X_n)^T$ be the test score vector, where X_i denotes test score of the *i*-th subject. Arranging the components of the vector $\boldsymbol{X}$ in decreasing order will give ranks of the individuals who took the test.

Consider the maximum possible test score vector $\mathbf{I}$ of order $n \times 1$ where $I_i = m \; \forall i = 1,2, \ldots \ldots m$. Difficulty and discriminating value of items and test can be obtained using cosine of the angle between the vectors $\mathbf{X}$ and $\mathbf{I}$ involving inner-product of the two vectors and length of the vectors.

By definition, if angle between two vectors **X** and **Y** is θ, then $Cos\theta = \frac{X.Y}{\|X\|\|Y\|}$ where$\|X\|$ denotes length of the vector **X** and is defined as $\|X\| = \sqrt{\sum_{i=1}^{n} X_i^2}$. $\|Y\|$ is defined accordingly. This gives the novel area of angular statistics where $Cos\theta$ gives similarity between two vectors of same dimension. Note that, for acute angle θ, $0 \leq Cos\theta \leq 1$; $Cos\theta = 0$ if and only if **X** and **Y**are orthogonal; $Cos\theta_{XY} = 1 \Leftrightarrow$**X** = **Y**. However, the triangle inequality is not satisfied i.e. $Cos\theta_{XY} + Cos\theta_{YZ} \geq Cos\theta_{XZ}$ is not always true for $\boldsymbol{X \neq Y \neq Z}$

2.1 Difficulty value of test:

Let $\emptyset$ be the angle between the vectors **X** and **I.** Here, $\|I\| = m\sqrt{n}$.

So,$Cos\emptyset = \frac{m\sum X_i}{\|X\| m\sqrt{n}} \Longrightarrow \bar{X} = \frac{\|X\| Cos\emptyset}{\sqrt{n}}$ (1.1)

Thus, test mean is equal to product of length of the score vector and cosine of the angle between the score vector and the maximum possible test score vector divided by square root of sample size.

From (1.1), $\bar{X}^2 = \frac{\|X\|^2 cos^2 \emptyset_X}{n}$

Now $Sin^2\emptyset = 1 - \frac{n\bar{X}^2}{\|X\|^2} \Longrightarrow \|X\|^2 Sin^2\emptyset = \|X\|^2 - n\bar{X}^2$

$\Longrightarrow$ Test variance $S_X^2 = \frac{\|X\|^2 Sin^2\emptyset}{n}$ (1.2)

Thus, SD of test score is product of the length of the test score vector and sine of the angle between the test score vector and the maximum possible test score vector divided by the sample size.

If **X** coincides with **I**, then the test is extremely easy since each subject has got maximum possible score. Difficulty value of a test should consider two fold criteria viz. $\emptyset$ and ratio of $\|X\|$and $\|I\|$. Accordingly, difficulty value of a test ($Diff_T$) may be expressed as :

$Diff_T = \frac{\|X\|}{\|I\|}\cos\emptyset = \frac{\bar{X}}{m}$ (1.3)

Note that $0 \leq Diff_T \leq 1$ and higher the value of $Diff_T$, easier is the test.

(1.3) defines difficulty value of a test as a ratio of length of the observed score vector and length of the idle vector, multiplied by cosine of the angle between the two vectors, keeping harmony with the usual notion of difficulty value of a test which actually measures degree of easiness of a test.

2.2 Difficulty value of item:

Difficulty value of an item can also be found in line with (1.3). Here, components of *n*- dimensional item score vector are zeros and ones. Let $\boldsymbol{I_i}$ be the maximum possible score vector for an item where each component is equal to 1. If *k* – persons ($k \leq n$) answer the *i*-th item correctly, then $\|X_i\| = \sqrt{k}$, $\|I_i\| = \sqrt{n}$ and $\cos\emptyset_i = \sqrt{\frac{k}{n}}$.

Thus, difficulty value of the *i*-th item ($Diff_i$) can be expressed as $Diff_i = Cos^2\ \emptyset_i = \frac{k}{n}$ (1.4)

Clearly, $0 \leq Diff_i \leq 1$

It may be observed that difficulty value of an item as per (1.4) coincides with normal idea of proportion of persons passing an item and can be taken as empirical probability of passing an item. $Diff_i$increases monotonically with increase in *k*. The curve of $Diff_i$ is a positively slopped.

The approach also helps to find difficulty value of a test in terms of item difficulty values.

Now $\bar{X} = \frac{\sum_{i=1}^{m} k_i}{n} = \sum_{i=1}^{m} Diff_i$

Thus, from (1.3), $Diff_T = \frac{\sum_{i=1}^{m} Diff_i}{m}$ (1.5)

(1.5) expresses difficulty value of the test in terms of item difficulty values.

2.3 Discriminating value of test:

If the vector **X** makes a zero degree angle with the vector **I**, then the test fails to discriminate the subjects. So ∅ or some function of ∅ will reflect the discriminating value of a test. Since it is desirable for the discriminating value to lie in [0, 1], *tan* ∅ will measure the discriminating value of a test. Thus,

$Disc_T = tan\emptyset = \frac{S_X}{\bar{X}}$ [From (1.1) and (1.2)] (1.6)

where $Disc_T$ denotes the discriminating value of a test.

Thus, discriminating value of a test is the ratio of SD and mean of the test score i.e. Coefficient of variation (CV) of the test scores.

2.4 Discriminating value of item:

Discriminating value of an item can be similarly defined by $Disc_i = \frac{S_{X_i}}{\bar{X}_i}$ (1.7)

where $Disc_i$ the discriminating is value of the *i*-th item; $\bar{X}_i$ is the mean score of the *i*-th item and S_{X_i} is the SD of the *i*-th item.

For the *i*–th item, the components of vector $\mathbf{X_i}$ are *k*- numbers of one's and rest zeros, if

k< n persons could answer the item correctly. Thus, score of the *i*–th item can be taken as a Binomial variate with parameters n and p where p is the probability of correct answer and is equal to $Diff_i = \frac{k}{n}$. Mean and SD are np and $\sqrt{npq}$ respectively, where $q = 1 - p = \frac{n-k}{n}$

Thus, co-efficient of variation of the *i*-th item (CV_i) is $\frac{\sqrt{npq}}{np} = \sqrt{\frac{q}{np}} = \sqrt{\frac{q}{k}} = \sqrt{\frac{n-k}{nk}}$

Thus, $Disc_i = \frac{S_{X_i}}{\bar{X}_i} = \sqrt{\frac{n-k}{nk}}$ (1.8)

Clearly, $Disc_i \geq 0$

The equation (1.8) avoids usual range of item discrimination values between - 1.0 to 1.0.

$Disc_i$ decreases with increase in *k.* Thus, $Disc_i$ curve is negatively sloped.

The $Disc_i$ may be multiplied by 100 and call it Percentage discriminating value of the item. Thus, Percentage discriminating value of the *i*-th item = $100.Disc_i$ (1.9)

2.5 Relationship between $Disc_i$ and $Diff_i$:

From (1.8), ${Disc_i}^2 = \frac{n-k}{nk}$.

Putting $k = n.Diff_i$ from (1.4), we get ${Disc_i}^2 = \frac{1-Diff_i}{n.Diff_i} = \frac{1-Diff_i}{k}$ (1.10)

i.e. square of discriminating value of an item is equal to (1-difficulty value of the item) divided by number of correct response (*k*) to the item.

It may be noted that low number of correct response (*k*) to the item means lower item difficulty value which implies higher discriminating value of the item, as per equation (1.10). Thus, relationship between $Diff_i$ and $Disc_i$ will be negative.

Observations:

(i) The discriminating value of an item is equal to the ratio of SD and mean of the item score i.e. coefficient of variation of the item score (CV_i)

(ii) If *k*= 0 i.e. the item is so difficult and no subjects could pass the item, then Discriminating value is not defined for the item. Clearly, such items with zero mean or infinite $Disc_i$ to be rejected without further investigation.

(iii) If *k* = *n* i.e. if all the subjects pass an item, then discriminatory value is zero for that item.

(iv) $Disc_i = 1$ implies $k = \frac{n}{n+1}$ which is a fraction. Thus, $0 \leq Disc_i < 1$, unlike the usual method using upper 27% and bottom 27% of data where discriminating index ranges between -1 to +1.

(v) Non-zero $Disc_i$ is maximum for k= 1 and minimum when k= (n-1). Thus, $Disc_i$ decreases monotonically with increase in k. In other words, the Percentage $Disc_i$ curve is negatively slopped non-linear curve. Equation (1.8) suggests that the curve showing $100.Disc_i$ and k has the form of a rectangular hyperbola for a given value of n.

(vi) Discriminating value of test and also item by CV has desired properties. Moreover, it is easy to estimate population CV as $\frac{\sigma}{\mu}$ where μ and $\sigma$are unbiased estimate of population mean and SD respectively.

(vii) If i-th and j-th items ($i \neq j$) have same mean, the item with lower SD will have lower CV and lower $Disc_i$.

(viii) To find value of k_0 for which $Disc_i = Diff_i$, one needs to solve the equation $\sqrt{\frac{n-k}{nk}} = \frac{k}{n}$ or

$$k^3 = n(n-k) \qquad (1.11)$$

In general, for a given value of n, value of k may be obtained through iterative solutions and choosing k appropriately between two successive integers between which k lies to satisfy (1.11) Alternately, k could be taken as the value (to the nearest integer) where the negatively slopped Percentage$Disc_i$ curve intersects with the positively slopped $Diff_i$ curve.Solution of (1.11) may be denoted as k_0.

Deletion of items:

Selection of items could be choosing the acceptance region as $(k_0 \pm \Delta)$ where Δ= 2SD ofdistribution of item difficulty values or item discriminating values. Choosing Δ= 3SD may result in discarding too few items and may not be desirable from the practical point of view.

In addition, considering skewness of distribution of $Diff_i$(or $Disc_i$), few more items having high concentration at the tail may be discarded.

However, choice of Δ may depend on original number of items in the test, type of test, whether to measure single dimension or multi dimensions and also considering relationship between test discrimination and test reliability.

2.6 Relationship between difficulty values and discriminating value of a test

From (1.3) and (1.6) we get $Diff_T . Disc_T = \frac{S_X}{m}$ (1.12)

i.e. product of difficulty value and discriminating value of a test is equal to SD of the test divided by number of items in the test. Discarding of few easy items (with high values of k) and few extremely difficult items (with very low values of k) will reduce m, and in turn may increase $Diff_T . Disc_T$.

2.7 Item – total correlation:

Point-biserial correlation coefficient (r_{pbs}) is the proper statistic to reflect item-total correlation i.e. the degree of relationship between score of an item (dichotomous variable) and test score (interval/ratio scale). r_{pbs} for the i-th item is defined as

$$r_{pbs(i)} = \frac{(M_{pi} - M_{qi})\sqrt{p_i q_i}}{S_X} \qquad (1.13)$$

where$r_{pbs(i)}$: Point-biserial correlation coefficient for the i-th item

M_{pi}:Test mean for persons answering the i-th item correctly (i.e., those coded as 1s)

M_{qi}: Test mean for persons answering the i-th item incorrectly (i.e., those coded as 0s)

S_X: Standard deviation of the test scores

p_i: Proportion of persons answering correctly i-th item = $\frac{k_i}{n}$ where score of the i-th item is k_i

q_i =1- p_i

Note that

i) $M_{pi} + M_{qi} = \bar{X}$(Test mean). Thus, $M_{pi} - M_{qi} = 2M_{pi} - \bar{X} = \bar{X} - 2M_{qi}$

ii) $p_i = \frac{k}{n} = Diff_i$ by (1.4).

iii) q_i =1- p_i = 1- $\frac{k}{n} = \frac{n-k}{n} = k.Disc_i^2$by (1.7) and (1.8)

iv) $S_X = \bar{X}.Disc_T$by (1.6)

Putting the above in (1.13) and considering that *k*-persons could answer the *i*-th item correctly, we get

$r_{pbs(i)} = \frac{(M_{pi}-M_{qi})\sqrt{\frac{k_i}{n}(\frac{n-k_i}{n})}}{\bar{X}.Disc_T}$ from (1.4) and (1.6)

$= \frac{(M_{pi}-M_{qi})\sqrt{Diff_i(1-Diff_i)}}{\bar{X}Disc_T}$ (1.14)

(1.14) depicts a negative relationship between item-total correlation, in terms of point biserial correlation and discriminating value of the test. High value of r_{pbi}indicates that persons who correctly answered the *i*-th item have done well overall on the test. Thus, r_{pbi}could be taken as measure item reliability. Clearly, $r_{pbi} \geq 0$ if $(M_{pi} \geq M_{qi})$

(1.14) can be further simplified as

$r_{pbs(i)} = \frac{(M_{pi}-M_{qi})\sqrt{Diff_i(\frac{Disc_i^2}{k_i})}}{m.Diff_T.Disc_T}$ from (1.10)

$= \frac{(M_{pi}-M_{qi})Disc_i\sqrt{Diff_i}}{k_i.m.Diff_T.Disc_T} = \frac{(M_{pi}-M_{qi})}{k_i.m}.\frac{Disc_i}{Disc_T}\frac{\sqrt{Diff_i}}{Diff_T}$

Now, $Disc_i.\sqrt{Diff_i} = \sqrt{\frac{(1-Diff_i)(Diff_i)}{k_i}} = \frac{\sqrt{n-k_i}}{n}$ and $Diff_T.Disc_T = \frac{S_X}{m}$

Thus, $r_{pbs(i)} = \frac{(M_{pi}-M_{qi})}{k_i.m}.\frac{\sqrt{n-k_i}}{n}.\frac{m}{S_X} = \frac{(M_{pi}-M_{qi})\sqrt{n-k_i}}{nk_iS_X}$ (1.15)

(1.15) may be taken as computational formula for $r_{pbs(i)}$

While reliability is a measure of association or similarities of two vectors, discrimination is a measure of dissimilarities or distance between the vectors. Higher values of item reliability are desirable. If two items have equal item discriminating value, then the item with higher variance is preferred to be retained

2.8 Relationship between test reliability and discriminating value of a test:

Let k_i be the number of persons answering the *i*-th item correctly. Then, from (1.7), variance of the *i*-th item $S_{X_i}^2 = \bar{X_i}^2.Disc_i^2$. Thus, sum of item variances,$\sum_{i=1}^m S_{X_i}^2 = \sum_{i=1}^m \bar{X_i}^2.Disc_i^2$

From (1.6), test variance $S_X^2 = \bar{X}^2.Disc_T{}^2$.Thus, test reliability in terms of Cronbach alpha is $\alpha = (\frac{m}{m-1})(1 - \frac{\sum_{i=1}^m \bar{X_i}^2.Disc_i^2}{\bar{X}^2.Disc_T{}^2})$ (1.16)

(1.16) expresses α in terms of terms of discriminating value of items and discriminating value of the test. For uni-dimensional test, impact of deletion of an item on alpha can be worked out finding changes in respective item and test parameters.

However, in general, considering theoretical definition of test reliability, $r_{tt} = \frac{S_T^2}{S_X^2}$, we get

$r_{tt}(Disc_T)^2 = \frac{S_T^2}{\bar{X}^2} = (\frac{S_T}{\bar{X}})^2 = (\frac{S_T}{\bar{T}})^2$ (1.17)

Thus, product of test reliability and square of test discriminating value is equal to square of CV of true scores. However, verification of the relationship may require finding test reliability as per the definition

Chakrabartty, (2013) proposed a method of obtaining test reliability as per the theoretical definition along with computation of error variance and true score variance from single administration of the test. The method involves an algorithm for splitting the test in two parallel halves with almost equal mean and SD and using lengths of score

vector of each such sub-test and the angle between the two vectors representing scores of the two parallel sub-tests. Method of obtaining test reliability as per theoretical definition is briefly discussed here.

If a test administered among n-subjects is dichotomized in two parallel halves say g-th and h-th sub-tests, two points $\boldsymbol{X_g}$and $\boldsymbol{X_h}$are obtained in the n -dimensional space. As per classical definition, two tests "g" and "h" are parallel if $T_{gi} = T_{hi}$and $S_{eg} = S_{eh}$, from which one can derive $\overline{X_g} = \overline{X_h}$ and $S_{Xg}^2 = S_{Xh}^2$.

Also, $X_g = T_g + E_g$and $X_h = T_h + E_h$. Now $T_{gi} = T_{hi} \Longrightarrow X_{gi} - X_{hi} = E_{gi} - E_{hi}$, so that

$$\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|Cos\theta_{gh} = \|E_g\|^2 + \|E_h\|^2 - 2\|E_g\|\|E_h\|Cos\theta_{gh}^{(E)}$$

whereθ_{gh} is the angle between $\boldsymbol{X_g}$and $\boldsymbol{X_h}$and $\theta_{gh}^{(E)}$ is the angle between $\boldsymbol{E_g}$and $\boldsymbol{E_h}$. But correlation between error scores of two parallel tests is zero. Thus,

$$\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|Cos\theta_{gh} = \|E_g\|^2 + \|E_h\|^2 = NS_E^2$$

since$S_E^2 = \frac{1}{N}\sum(E_{gi} + E_{hi})^2 = \frac{1}{N}[\|E_g\|^2 + \|E_h\|^2]$

The above equation suggests

$$S_E^2 = \frac{1}{N}[\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|Cos\theta_{gh}] \quad (1.18)$$

$$\text{Hence, } r_{tt} = 1 - \frac{\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|cos\theta_{gh}}{NS_X^2} \quad (1.19)$$

Equation (1.19) gives theoretical test reliability in addition to estimate error variance of the test by (1.18) and hence true score variance.

Equation (1.19) and (1.18) may also help to find impact of deletion of items on error variance of the test and test reliability respectively. Items must not be deleted if test reliability gets reduced or error variance of the test gets increased.

2.9 Deletion of items:

Based on point of intersection of the $Diff_i$ and $Disc_i$ curves:

Let $k_0$be the value for which $Diff_i = Disc_i$. Selection of items to increase discriminating value of the test could be choosing those items for which k lies in a neighborhood of k_0 like $(k_0 \pm \Delta)$ where Δ may be 2SD of distribution of k. Alternatively, considering values of item difficulty $(K_{0(Diff)})$ and/or item discriminating $(K_{0(Disc)})$ corresponding to k_0,the acceptance region could be $(K_{0(Diff)} \pm \Delta)$ or $(K_{0(Disc)} \pm \Delta)$ where Δ= 2SD ofdistribution of item difficulty values or item discriminating values. $(k_0 \pm 3SD)$ may result in discarding too few items and may not be desirable from the practical point of view.

However, choice of Δ may depend on original number of items in the test, type of test, whether to measure single dimension or multi dimensions and also considering relationship between test discrimination and test reliability.

Based on Item reliability:

Items with marginal point biserial correlation coefficient may be adjusted or removed.

Based on other criteria:

Considering skewness of distribution of $Diff_i$(or$Disc_i$), few more items having high concentration at the tail may be discarded.

Deletion of items is advisable only when reliability of the test improves upon deletion.

3. Empirical verification:

Data: A Selection Test was administered to 911 candidates. The test had 50 items and maximum time given was 90 minutes. Scores of those 911 candidates were considered for empirical verification of the foregoing method. Here, n= 911 and m = 50

For the test, Mean = 20.49506; Median = 10; Mode = 11 and Variance = 11.95799. Thus, the distribution of test score was not symmetric.

By (1.3), $Diff_T = \frac{\|X\|}{\|I\|}\cos\emptyset = \frac{\bar{X}}{m} = 0.409901 \Rightarrow$ Test was moderately difficult

By (1.6) $Disc_T = tan\emptyset = \frac{S_X}{\bar{X}} = 0.168725 \Rightarrow$ Test had rather poor discriminating value.

Item discriminating value as defined in (1.8) exceeded 0.1 only for one item (Item no.40 with k= 30. For meaningful comparative statements of item discriminating values and item difficulty values, each $Disc_i$ was multiplied by 100

Frequency distribution of Item difficulty values and Percentage Item discriminating values are shown in Table – 1

Table – 1. *Frequency distribution of $Diff_i$ and Percentage$Disc_i$*

Item difficulty values ($Diff_i$)		(Percentage $Disc_i$)
Intervals	Frequency	Frequency
Up to 0.1	1	16
0.1 – 0.2	4	12
0.2 – 0.3	13	8
0.3 – 0.4	7	6
0.4 – 0.5	8	4
0.5 – 0.6	7	2
0.6 – 0.7	6	--
0.7 – 0.8	4	1
0.8 – 0.9	--	--
0.9 and above	--	1
Total	50	50

Graphs showing difficulty values and discriminating values of items are given in Figure 1.

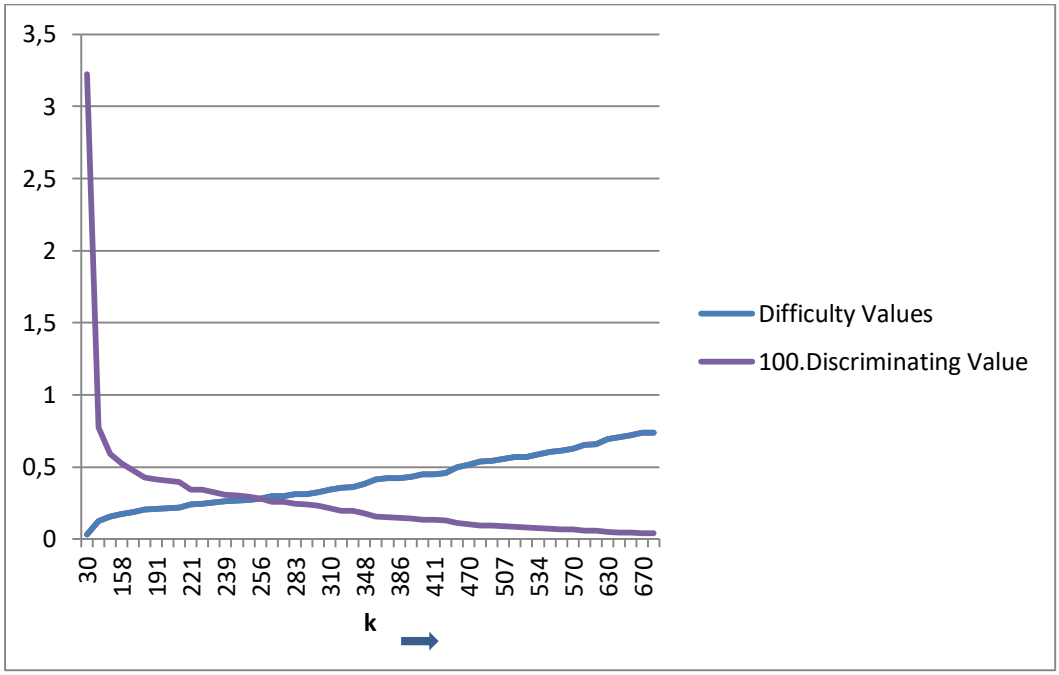

Figure. 1: *Item Difficulty Values and Percentage Discriminating Values of İtems*

Clearly, as k increases, item difficulty curve increases and percentage item discriminating curve decreases. The two curves cut at a point (k_0)= 368. Note that at k=368, item difficulty is 0.40395 and percentage item discriminating is 0.40245, the difference being 0.00149. Shifting k_0 to the right will increase proportion of items with high difficulty values (and low discriminating values). Thus, choice of k_0 may be considered while

deleting number of items from the test. In the instant case, the test had 25 items (i.e. 50% of items) in the left of k_0

The items to be ignored may be those lying outside ($k_0 \pm 2$SD of Item scores). Mean and SD of item scores, ($Diff_i$) and ($Disc_i$) along with acceptance regions are given in Table -2.

Table – 2. *Acceptance Region of Items*

	Mean	SD	Acceptance region ($k_0 \pm 2SD$)
Item score	20.49	167.5	368±335.18
$Diff_i$	0.02249	0.1839	0.02249 ±0.3679
$Disc_i$	0.00258	0.025	0.00258 ±0.05089

Each of the above method resulted in discarding the item no. 40 only with *k*=30, being extremely difficult i.e. lowest Diff. value (0.033) and highest Disc. value (0.17954).

Deletion of an item will change values of $Diff_T$ and $Disc_T$. For example, if the most difficult item with *k*=30 is deleted, new $Diff_T$ increased to 0.4175945 from original value of0.409901 and new $Disc_T$ got reduced to 0.1739181from original value of 0.168725.

The present data had 16 items with *k*-values less than 368 (k_0) and 34 items with *k*-values exceeding k_0 (rather easy items). Easy items with high *k*-values (i.e. high Diff. values say ≥ 0.70 implyinglow Disc. values≤ 0.022) may also be considered for discarding. Adoption of this criteria implies discarding of additional four items (viz. Item no. 1(*k*= 670*);* 8 (*k*= 654); 33 (*k*= 645*)* and 44(*k*= 672).

Correlation between difficulty values and discriminating values of items:

The graph of item difficulty values and item discriminating values suggest that the $r_{Diff_iDisc_i}$be negative. Value of $r_{Diff_iDisc_i}$was (-) 0.579586.

Test reliability:

The 50 items of the test were dichotomized to *g*-th and *h*-th subtests following the procedure given by Chakrabartty (2013). Resultant parallel halves as per the proposed iterative method are given in Table-3

Table–3. *Splitting as per the iterative process*

***g*-th sub-test**		***h*-th sub-test**		**Difference (*g*-*h*)**
Item No.	**Item Score**	**Item No.**	**Item Score**	
41	113	40	30	83
5	158	16	143	15
7	171	43	187	-16
20	194	28	191	3
47	197	30	221	-24
49	230	15	222	8
19	239	17	243	-4
11	256	26	248	8
21	273	27	273	0
2	283	18	285	-2
10	294	32	310	-16
50	328	6	325	3
23	348	42	375	-27
29	386	48	385	1
13	393	22	410	-17
14	417	3	411	6
38	452	34	470	-18
39	491	12	493	-2
36	507	4	519	-12
24	534	25	520	14
46	551	45	558	-7
35	595	9	570	25

37	601	31	630	-29
8	654	33	645	9
1	670	44	672	-2
Sum	9335		9336	-1
Mean	10.25		10.25	0
SD	66.70		67.11	-0.41

Observations:

- Splitting the test by the iterative process resulted in $\overline{X_g} = \overline{X_h}$ = 10.25 and $|S_g - S_h|$= 0.418.

Marginal difference (0.418) between the SDs of the g-th and h-th tests (much less than the same obtained from odd–even split half). Accordingly, splitting half as per the iterative process was considered better for almost equality of means and SDs.

- Split-half reliability r_{gh} as the correlation between person scores in the g-th and h-th subtests was found to be 0.380813

Here, $\|X_g\| = 315.6169$; $\|X_h\| = 315.6058$ and $Cos\theta_{gh} = 0.975479$

Theoretical reliability of the test $r_{tt} = 1 - \frac{\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|cos\theta_{gh}}{n.S_X^2}$ = 0.551577 and

$S_E^2 = \frac{1}{n}[\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|Cos\theta_{gh}]$ = 5.362239

and true score variance = 6.595749

Theoretical reliability of the test was higher than the Split-half reliability but lower than Cronbach α(0.78)

4. Findings and Conclusions

New measures of difficulty and discriminating values of binary items and test consisting of such items were proposed considering cosine similarity i.e. length of two score vectors and angle between them. The measures considered entire data and not only top 27% and bottom 27% of data. Difficulty value of a test ($Diff_T$) is defined as the ratio of length of the observed score vector and length of the idle vector, multiplied by cos Øwhere Øis the angle between the two vectors, keeping harmony with the usual notion of difficulty value of a test which actually measures degree of easiness of a test. Discriminating value of a test ($Disc_T$) is $tan\emptyset$ which is the ratio of SD and mean of the test score. Similarly, discriminating value of an item ($Disc_i$) is equal to the ratio of SD and mean of the item score i.e. coefficient of variation (CV). Here, $0 \leq Disc_T \leq 1$ and similar inequalities hold for $Disc_T, Diff_T$and $Diff_i$. Discriminating value of test and also item in terms of CV has desired properties. Moreover, it is easy to estimate population CV.

Relationship derived between item difficulty value and item discriminating values. As number of correct answer to an item (*k)* increases, item difficulty curve increases and item discriminating curve decreases. The point of intersection of the two curves (k_0) is a data driven criterion which may be considered in deciding the items to be deleted which are lying outside acceptance region defined as an interval ($k_0 \pm \Delta$) where Δ could be taken as 2SD of distribution of Item scores or $Diff_i$ *or* $Disc_i$. Other criteria of item deletion could be based on skewness of distribution of item scores or item reliability i.e. point biserial correlation coefficient of an item. However, actual deletion of items needs to consider impact of such deletions on reliability of the test.

Relationship established between difficulty value and discriminating value of a test. Cronbach alpha was expressed and computed using sum of item difficulty values and test discriminating value. Similarly, relationship derived between test discriminating value and test reliability as per theoretical definition. Further, relationship derived between item reliability to depict Item-total correlations, in terms of Point biserial correlation ($r_{pbs(i)}$) with test parameters like $Disc_T, Diff_T$ and also item parameters like $Diff_i$ and $Disc_i$. In fact, $r_{pbs(i)}$has a negative relationship with test discriminating value and number of items in the test.

Test reliability as per theoretical definition was computed. Empirically, value of theoretically defined reliability was greater than the split-half correlation but marginally lower than Cronbach alpha. Future investigations may be undertaken to verify the proposed measures and their factors with multiple data sets.

Acknowledgement: Nil

Funding details: did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interests: Nil

5. References

Chakrabartty, Satyendra Nath (2013). Best split-half and maximum reliability. *IOSR Journal of Research & Method in Education*, 3(1), 01-08.

Chauhan,P.R., Ratrhod, S. P., Chauhan, B. R., Chauhan, G. R., Adhvaryu, A. and Chauhan, A.P. (2013) Study of difficulty level and discriminating ındex of stem type multiple choice questions of anatomy in rajkot, *BIOMIRROR, 4*(06), 1-4.

Denga, I. (2009).*Educational measurement, continuous assessment and psychological testing*. Rapid Educational Publishers.

Ebel, R.L. and Frisbie, D.A. (1991) *Essentials of educational measurement*. Prentice Hall of India Pvt. Ltd.

Ferrando, Pere. J. (2012).Assessing the discriminating power of item and test scores in the linear factor-analysis model. *Psicologica*, 33, pp. 111-134.

Hankins M. (2007). Questionnaire discrimination: (re)-introducing coefficient Delta. *BMC Medical Research Methodology*. 7:19. doi: 10.1186/1471-2288-7-19.

Henrysson, S. (1971) Gathering, analyzing and using data on test items. In R. L. Thondike (Ed.) *Educational measurement* (2nd ed. pp 130-159). American Council on Education.

Kehoe, Jerard (1995) *Basic ıtem analysis for multiple-choice tests.* ERIC/AE Digest. https://pareonline.net/getvn.asp?v=4&n=10

McDonald, R.P. (1999) *Test theory: A unified treatment*. Lawrence Earlbaum Associates, Inc.

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

Rao C, Kishan Prasad H L, Sajitha K, Permi H, Shetty J. (2016) Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *Int J Educ Psychol Res,*2:201-4

Rudner, Lawrence M and Schafes, William (2002) Reliability: *ERIC Digest*. www.ericdigest.org/2002-2/reliability/htm

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, Si–Mui and Rasiah, R. I.(2006) Relationship between ıtem difficulty and discrimination ındices in true/false type multiple choice questions of a para-clinical multidisciplinary paper, *Annals of the Academy of Medicine,* Singapore, *35*(2), 67-71.

Tzuriel, D. and M. Samuels. (2000). Dynamic assessment of learning potential: Inter-rater reliability of deficient cognitive functions, type of mediation and non-intellective factors. *Journal of Cognitive Education and Psychology*, 1, 41-64.