Journal of Modern Applied Statistical Methods

Volume 23 | Issue 1 Article 13

Relationships among Quality Parameters of Tests

Satyendra Nath Chakrabartty
Indian Statistical Institute, Indian Institute of Social Welfare and Business Management, Indian Ports Association, India, Email: chakrabarttysatyendra3139@gmail.com, ORCID: 0000-0002-7687-5044

Recommended Citation

Satyendra Nath Chakrabartty (2024). Relationships among Quality Parameters of Tests. Journal of Modern Applied Statistical Methods, 23(1), https://doi.org/10.56801/Jmasm.V23.i1.13

Journal of Modern Applied Statistical Methods
2024, Vol. 23, No. 1,
Doi: 10.56801/Jmasm.V23.i1.13

ISSN 1538 – 9472

Relationships among Quality Parameters of Tests

Satyendra Nath Chakrabartty
Indian Statistical Institute, Indian
Institute of Social Welfare and Business
Management, Indian Ports Association,
India

Background: Tests involve distinct test parameters like reliability, validity, difficulty value, discriminating value, responsiveness, etc. To assess quality of test, relationships among the test parameters are needed.
Method: The paper describes methods of finding discriminating value of test/scale ($Disc_T$) and item ($Disc_i$) considering the entire data and derives relationship between theoretically defend test reliability ($r_{tt-theoretical}$) and $Disc_T$, and also with factorial validity (FV). Relationship also derived between FV and α_{PCA} both in terms of eigenvalues. Transformation of discrete item scores to continuous scores $P_i \sim N(\mu_i, \sigma_i)$ and test scores (P-scores) as $\sum P_i$ enabling better arithmetic aggregation, meaningful comparisons and satisfying desired properties discussed.
Findings: Relationships between $r_{tt-theoretical}$ and $Disc_T$ and also with FV was non-linear. Thus, one cannot increase all the parameters simultaneously. However, FV of standardized scores ($FV_{Z-scores}$) and α_{PCA} was positively related. P-scores facilitate parametric analysis like PCA, testing and estimation of mean, variance of item, test, coefficient of variation, Cronbach alpha at population level, FV in terms of eigen values, Reliability byα_{PCA}, finding equivalent scores of two or more instruments, assessment and testing significance of responsiveness, testing of $H_0: r_{tt} = 1$, etc. Significance of FV can be tested by Tracy–Widom distribution.
Conclusions: Scores emerging from MCQ type tests or scales/questionnaire containing items with different number of response-categories may be transformed to follow normal distribution parameters of which can be estimated. Derived relationships of test reliability with discriminating value or with factorial validity have potentials to find optimal value of one parameter to maximize another parameter.

Keywords: Discriminating value, Eigenvalue, Factorial validity, Normal distribution, Reliability.

1. Introduction

Measurements in tests or scales often involve several distinct test parameters like reliability, validity, difficulty value, discriminating value, responsiveness, etc. Scales are instruments to measure constructs in affective domain and cognitive features like knowledge and understanding of concepts and topics. To assess quality of test as a whole, it is needed to consider relationships among the test parameters. A test must be valid (high value of validity) with high precision (high value of reliability), able to discriminate the subjects taking the test (high discriminating value) and able to assess changes with time (responsiveness). A test shows different combinations of the above said parameters. For example, a test could be highly reliable with poor discriminating power and of moderate validity. The same is true for the items included in the test.

Discriminating value of a test $(Disc_{Test})$ is directly related to quality of test scores as a measure of the trait (McDonald, 1999). Test reliability or test validity does not indicate the degree of discrimination offered by a test (Hankins, 2007). Inclusions of items with negative or zero discrimination value or poor item reliability are highly undesirable since they result in measurement disturbances and reduce test qualities. Item discriminating values $(Disc_i)$ are usually lower for non-homogeneous tests. $Disc_{Test}$ ranging between - 1.0 to 1.0 (Denga, 2009) indicates 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. In addition, several measures are used to find goodness of the items. Examples include item reliability by item-total correlations (with or without that item) (Tzuriel and Samuels, 2000), bi-serial correlation between an item score and test scores of all subjects (r_{bs}) (Ebel and Frisbie, 1991), point bi-serial correlation (r_{pbs}) reflecting predictive validity of the test (Henrysson, 1971), Spearman's rank correlation, etc. However, r_{bs} tends to favor items of average difficulty. Researchers tend to differ on cut-off 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 with $r_{pbs} \leq 0.19$.

Responsiveness is the ability of a tool to accurately detect changes in the purported construct (s) across time. Naturally it involves longitudinal data to assess changes in score of one or a group of individuals. Mokkink et al. (2021) considered responsiveness as longitudinal validity, as it relates to the degree to which an instrument is able to measure change in the construct to be measured. Other relevant issue is finding the equivalent score of test-2 for a given score of test-1 where the two tests have different length (number of items), width (number of response-categories) and distribution of scores.

Need is felt to have reliable method of computing quality parameters of test and items and find their relationships under classical test theory (CTT). Model based complex Item Response Theory (IRT) was not considered primarily for its requirement of large sample size, testing whether the data fits the model and number of strict assumptions including a curvilinear relationship between item score and construct score against a simple linear relationship between them by CTT.

The paper derives relationships of test reliability as per definition (ratio of true score variance and observed score variance) with factorial validity, and difficulty/discriminating value of items and tests without sacrificing any portion of data. Transformation of discrete item scores to continuous scores $P_i \sim N(\mu_i, \sigma_i)$ and test scores (P-scores) as $\sum P_i$ enabling better arithmetic aggregation, meaningful comparisons and satisfying desired properties discussed. Thus, the approach is an improvement over observation made by Rudner and Schafes, (2002) who mentioned that it is impossible to calculate reliability as per theoretical definition since true scores of individuals taking the test are unknown.

2. Literature survey

2.1 Reliability:

Test reliability (r_{tt}) is defined as ratio of true score variance (S_T^2) and observed score variance

(S_X^2) i.e. $r_{tt} = \frac{S_T^2}{S_X^2}$ (1)

However, in practice, test reliability is found by different approaches ignoring the definition. Popular methods of reporting test reliability requiring a single administration are: Cronbach alpha, Inter-item reliability, Split-half reliability for internal consistency to assess how well different items measure the same characteristic (Utwin, 1995). Test-retest reliability requiring two administrations is focused on stability. Inter-rater reliability indicates degree of agreement among the raters who rate, code, or assess the same phenomenon independently. Different approaches, based on different set of assumptions, result in different values of reliability and require different interpretations.

Test-retest reliability ($r_{Test-retet}$) is the correlation of scores emerging from two administrations of the test on the same group of individuals under the same conditions with a time gap, assuming true scores of an individual remain unchanged during the time gap. However, factors like practice effect, learning effect during the time gap, errors in the testing situations, etc. can influence $r_{Test-retet}$ values depending on time gap, for which there is no consensus. Thus, the assumption of unchanged true scores may not hold in reality. Focus could be to measure similarity of ranks of individuals or agreement of individual scores in two administrations. Interpretation of $r_{Test-retet}$ could be problematic since correlation is different from agreement. Clearly, $r_{Test-retet}$ is a necessary but not sufficient condition to demonstrate agreements. Berchtold (2016) preferred correlation than agreement. Jelenchick et al. (2012) used correlation, and not agreement to find $r_{Test-retet}$ of Internet Addiction Test (IAT) developed by Young (1998). Average IAT scores in 2nd administration decreased from the 1st administration and t-statistic was - 6.34

with 677 df, $p < 0.001$. Negative value of t-statistic is an evidence against the $H_0: \mu_{1st\ administration} = \mu_{2nd\ administration}$. Shifting from theoretical reliability as in (1), debate on stability and agreement continues for $r_{Test-retet}$

Parallel test or split-half reliability considers correlation of two parallel sub-tests resulting from dichotomization of the test scores. Split-half reliability depends heavily on method of dichotomization. Different splits give different values of split-half reliability. In addition, test of parallelism of the sub-tests are required.

Inter-item reliability takes average of inter-item correlations for assessment of extent to which items on a test/scale are assessing the same content (Cohen & Swerdlik, 2005). Such average depends significantly on magnitude and direction of correlations. Intuitively, average of $r_{xy} = 1$ and $r_{pq} = -1$ is zero. However, correlations are not additive (Garcia, 2012). Average of item-total correlations was disfavoured (Field, 2003). Fisher's Z-transformation $Z_r = \frac{1}{2}\ log_e[\frac{1+r}{1-r}]$ following Normal distribution may be undertaken for meaningful average of correlations. Taking inverse function of Z_r one can get $\bar{r}_n = \frac{e^{2\bar{Z}n-1}}{e^{2\bar{Z}n+1}}$ which accounts for sign of the correlations (Bewick et al. 2003). However, Fisher's Z-transformation to correlations, violating bivariate normality, may give spurious results (Zimmerman et al. 2003).

Alternatively, consider vectors $\boldsymbol{X}_{n\times1}$ and $\boldsymbol{Y}_{n\times1}$for variables X and Y; find deviation score vectors **x** and **y** where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$. Let θ be the angle between the vectors **x** and **y**, then $r_{XY} = Cos\theta_{xy}$=$\frac{x^Ty}{\|x\|\|y\|}$ where length of the vector **x** and **y** are $\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}$; $\|Y\| = \sqrt{\sum_{i=1}^{n} y_i^2}$ respectively and x^Ty denotes dot product of x and **y** = $\sum_{i=1}^{n} x_i y_i$. However, the triangle inequality law is not satisfied by $Cos\theta_{xy}$ i.e. $Cos\theta_{XY} + Cos\theta_{YZ} \geq Cos\theta_{XZ}$ is not always true for $\boldsymbol{X \neq Y \neq Z}$**.** Normalizing k-vectors to unit length, Rao (1973) gave a method of finding mean and dispersion of angles $\emptyset_1, \emptyset_2, \ldots\ldots\ldots, \emptyset_k$, as mean or most preferred direction is $\bar{\emptyset} = Cot^{-1}\frac{\sum_{i=1}^{k} Cos\emptyset_i}{\sum_{i=1}^{k} Sin\emptyset_i}$ and the dispersion = $\sqrt{1-r^2}$ where $r^2 = (\frac{\sum Cos\emptyset_i}{k})^2 + (\frac{\sum Sin\emptyset_i}{k})^2$.

Cos ($\bar{\emptyset}$) gives the average of $Cos\theta_{ij}'s$ i.e. average of $r_{XY}'s$

Cronbach alpha of an instrument with m-items is $\alpha = \frac{m}{m-1}\left[1 - \frac{Sum\ of\ item\ variance}{Test\ variance}\right]$.

Equivalently, alpha in terms of covariance is

$\alpha = \frac{m\ (Av.inter-item\ covariance\ among\ the\ items)}{Av.variance+(m-1)(Av.\ inter-item\ covariance\ among\ the\ items)}$ or following Cortina, (1993)

$$\alpha = m^2[\frac{mean\ inter-item\ covariance}{sum\ of\ all\ the\ elements\ in\ the\ variance-covariance\ matrix}]$$

Clearly, α increases with increase in m.

Cronbach alpha assumes all items are equivalent test units, test is one-dimensional, items are tau-equivalent i.e. all the factor loadings are same or equivalently the off-diagonal elements of the variance–covariance matrix of the component scores are same (Ogasawara, 2006). However, same factor loadings may not fit well to cognitive tasks, due to their designs and scoring algorithms (Pronk et al. 2022). If items are not essentially tau-equivalent, they measure different constructs, and alpha is underestimated. In addition, alpha requires uncorrelated errors and normality. However, many scales reports alpha despite finding several factors from Principal component analysis (PCA) or Factor analysis (FA). For example, ACT reports seven sub-scores and find internal scale score reliability for the sub-scores, averaged across five administrations (ACT, 1997). Huang and Tang (2013) computed eigenvalue and alpha for each of four factors (perceived usefulness (PU), perceived playfulness (PP), and resistance to change (RTC) of Mobile English learning satisfaction (MELS) and the construct with highest eigenvalue (6.027) had the maximum alpha (0.895). Use of alpha goes hand-in-hand with PCA since alpha has something to do with the factor structure of the test (De Hooge et al. 2007). Using results of PCA, Ten Berge and Hofstee, (1999) proposed test reliability.

$$\alpha_{PCA} = (\frac{m}{m-1})\ (\ 1 - \frac{1}{\lambda_1}) \qquad (2)$$

where λ_1 is the first (largest) eigenvalue of correlation matrix of m-number of items. Clearly, λ_1 gives maximum value of alpha for any linear combination of the m-items of the test. Mean value of alphas for factor scores of different subgroups were computed (Hampson et al. 2015). α_{PCA} for different dimensions like Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Intellect/Openness of Childhood Big Five personality measures were found (Edmonds et al., 2013). Thus, different methods of finding reliability may give different values of reliability even from the same data.

Several statistical techniques, estimation and testing of hypothesis, PCA, etc. assume normally distributed scores. In practice, normally distributed data are rare in education and social science. Chakrabartty (2022) transferred discrete item scores (X_i) to continuous equidistant scores (E_i) followed by standardization (Z_i) to follow $N(0,1)$ and further linear transformation to proposed scores (P_i) in a desired score range, say [1, 100] following$N(\mu_i, \sigma_i)$ and test/scale scores (P-scores) as sum of $P_i s$ i.e. convolution of the item scores which also follows normal distribution. $\sum P_i$ is more meaningful because of similarity of distribution of item scores and the resultant sum also follows normal distribution. Such P-scores satisfy desired properties like meaningful arithmetic aggregation, undertaking parametric statistical analysis like PCA, FA, ANOVA, statistical inferences like estimation and testing hypothesis of equality of means across time and space. For normally distributed test scores$P$, true score of an individual with $P = P_0$ is estimated by $P_0 \pm SEM$ where $SEM =$Sample S_E (Chakrabartty, 2022).

Kristof (1963) found distribution of alpha coefficient assuming that the off-diagonal elements of the variance-covariance matrix and that of the diagonal elements have the same values (known as compound symmetry) under normality. Normally distributed *P*-scores help to estimate variance of *i*-th item (σ_i^2) and test variance (σ_X^2) and estimate Cronbach alpha at population level as

$$\hat{\alpha} = \frac{m}{m-1}\left(1 - \frac{\sum_{i=1}^{m}\sigma_i^2}{\sigma_X^2}\right) \tag{3}$$

If the item scores are transformed to follow $N(0,1)$, $\hat{\alpha}$ = $\frac{m}{m-1}(1 - \frac{m}{\sigma_X^2})$ and estimation of σ_X^2 can be found from convolution of distributions of standardized item scores. In general, distribution of $\hat{\alpha}$ can be found using sample variance of *P*-scores following χ^2-distribution with (*n*-1) degrees of freedom(df); sum of χ^2 variables follows χ^2-distribution and $\frac{U/n_1}{V/n_2}$ ~ *F*-distribution with n_1 and n_2 df where $U \sim \chi^2_{n_1}$ and *V* $\sim \chi^2_{n_2}$(Larsen & Marx, 2010).

Chakrabartty (2022) suggested method to find test reliability (r_{tt}) as per its definition from single administration of the test. It involves dichotomizing a test to g-th and h-th sub-tests each containing $^{m}/_{2}$ -items where the subtests are parallel i.e. true score of *i*-th person in two subtests are equal ($T_{ig} = T_{ih}$) and error SD of the sub-tests are equal ($S_{eg} = S_{eh}$). Consider sub-tests scores as n-dimensional vectors X_g and X_h with length $\|X_g\|$ and $\|X_h\|$ where $\|X_g\| = \sqrt{\sum_{i=1}^{m/2} {X_{ig}}^2}$ and $\|X_h\| = \sqrt{\sum_{i=1}^{m/2} {X_{ih}}^2}$. Let θ_{gh} be the angle between X_g and X_h.

Here, $X_{ig} - X_{ih} = E_{ig} - E_{ih}$

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

where $\theta_{gh}^{(E)}$ is the angle between the vectors E_g and E_h

$\Rightarrow \|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|Cos\theta_{gh} = \|E_g\|^2 + \|E_h\|^2 = n{S_E}^2$ since correlation between error scores of two parallel tests is zero and ${S_E}^2 = \frac{1}{n}(E_{ig} + E_{ih})^2 = \frac{1}{n}(\|E_g\|^2 + \|E_h\|^2)$

$$\Rightarrow {S_E}^2 = \frac{1}{n}(\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|Cos\theta_{gh}) \tag{4}$$

$\Rightarrow {S_T}^2 = {S_X}^2 - {S_E}^2$ and

$$r_{tt} = 1 - \frac{\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|Cos\theta_{gh}}{n{S_X}^2} \tag{5}$$

Equation (5) gives test reliability and (4) gives value of error variance (square of standard error of measurement)

Considering $\|X_g\|$ = $\|X_h\|$ as they are parallel, equation (4) and (5) can be simplified as

$$ {S_E}^2 = \frac{2\|X_g\|^2(1 - Cos\theta_{gh})}{n} \tag{6}$$

and

$$r_{tt} = 1 - \frac{2\|X_g\|^2(1-Cos\theta_{gh})}{nS_X{}^2} \tag{7}$$

Theoretical reliability $r_{tt-theoretical}$ as defined above is different from split-half reliability(r_{gh}). For a selection test with *m*= 50 and *n*= 911, Chakrabartty (2021) found $r_{tt-theoretical} = 0.56$ which exceeded the Split-half reliability (r_{gh} =0.38) but was lower than Cronbach α (0.78).

Transformation of item and test scores to normally distributed *P*-scores before dichotomization, helps to test $H_0: r_{tt} = 1$ which is equivalent to H_0: $\sigma_X^2 = \sigma_T^2$ by *F*-test using test statistic $F = \frac{S_X^2}{S_T^2}$ and reject H_0 if $F > F_{\alpha,(n-1,n-1)}$. Confidence interval of $r_{tt} = \frac{\sigma_T^2}{\sigma_X^2}$ is given by $\frac{S_T^2/S_X^2}{F_{\alpha/2}} \leq \frac{\sigma_T^2}{\sigma_X^2} \leq \frac{S_T^2/S_X^2}{F_{1-(\alpha/2)}}$. In addition, testing H_0: $\bar{\mu}_g = \bar{\mu}_g$by *t*-test and H_0: $\sigma_{Xg}^2 = \sigma_{Xg}^2$ by *F*-test help to test whether *g*-th and *h*-th sub-tests are parallel. Other tests to show g-th and *h*-th sub-tests are parallel could be testing equality of regression lines $X = \alpha_1 + \beta_1 X_g$ and $X = \alpha_2 + \beta_2 X_h$ by ANOVA or by Mahalanobis $D^2 = d^T S^{-1} d$ where $d_i = \overline{X_{g\imath}} - \overline{X_{h\imath}}$ for the *i*-th item.

The approach of finding reliability as per definition can be extended to find reliability of a battery of tests to measure a finite number of constructs. After administration of the battery consisting of K-tests to *n*-individuals, values of S_X^2, S_E^2, S_T^2 and r_{tt} for each constituent test can be computed. If battery score is taken as sum of score of K-tests, battery reliability is

$$r_{tt(battery)} = \frac{\sum_{i=1}^{K} r_{tt(i)} S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2Cov(X_i, X_j)}{\sum_{i=1}^{K} S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2Cov(X_i, X_j)} \tag{8}$$

If $W_1, W_2, \ldots \ldots . . W_K$ are the weights to K-constituent tests of a battery where $W_i \geq 0 \ \forall \ i = 1,2,3, \ldots . K$ and $\sum W_i = 1$ and the battery score be $Y_i = \sum_{i=1}^{K} W_i X_i$. Here, $var(Y) = \sum_{i=1}^{K} W_i^2 var(X_i)$ and the battery reliability is

$$r_{tt(battery)} = \frac{\sum_{i=1}^{K} r_{tt(i)} W_i^2 S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2W_i W_j Cov(X_i, X_j)}{\sum_{i=1}^{K} W_i^2 S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2W_i W_j Cov(X_i, X_j)} \tag{9}$$

2.2 Validity:

For a measurement to be valid, it has to be reliable. While validity is associated with accuracy, reliability is about consistency. Therefore, an unreliable measurement cannot be valid. However, a measurement can be reliable without being valid.

Test validity is often measured by correlating test score (X) with a criterion score (Y). If $r_{XY} = 0.65$ (say), then 0.65 is the validity of X and also of Y. If r_{XY} is still more, test (X) may not be required. Moreover, r_{XY} could be influenced by factors like dis-similarity in dimensions covered, factor structures of X and Y, different score ranges, sample heterogeneity, etc. For a selection test, homogeneity of the selected individuals may lower the validity. For example, $r_{XY}$was - 0.93302 for $0 \leq X \leq 3.9$ and X ~*N (0, 1)* and $Y = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}X^2}$. However, for$-3.9 \leq X \leq 3.9$, r_{XY}= 0.00036.

Thus, truncated values of one or more variables (or homogeneity of data) may distort correlation between two variables. In other words, truncated score can underestimate or overestimate the validity. Validity of parametric analyses of scales generating Likert-type data is often unclear (Lantz, 2013).

Better could be to find test validity from a single administration considering scores of all individuals who took the test by factorial validity (FV) expressed as

$$\text{FV}= \frac{\lambda_1}{\sum \lambda_i} \tag{10}$$

where λ_1 is the highest eigenvalue corresponding to the main factor for which the scale was developed. Clearly, FV will be high for unidimensional tests giving high value of λ_1. Factorial validity reflects validity of the main factor for which the test was developed and accounts for $\frac{\lambda_1}{\sum \lambda_i} \times 100$ percent of overall variability. Such factorial validity from single administration of a test avoids the problems of construct validity and is independent of criterion scale (Parkerson et al. 2013).

Normally distributed P-scores enables undertaking of PCA and computations of λ_i's and component loadings of items (the eigenvector $\times \sqrt{\text{the eigenvalue}}$) which can be interpreted as the correlation of the item with the principal component or item validity. In addition, factorial validity is simple to comprehend; Item validity is given in terms of component loading where sum of item validities ≠ Scale validity.

In the context of Statistical Physics and Signal processing, Nadler (2011) found that $U = \frac{\lambda_1}{\sum \lambda_i} = \frac{\lambda_1}{T}$ follows a Tracy–Widom (TW) distribution, where T=$\sum \lambda_i$ is the trace of the variance-covariance matrix = Sum of item variances. Tracy–Widom (TW) distribution is a probability distribution of the normalized largest eigenvalue of a random Hermitian matrix whose entries are independently Gaussian-distributed. Data analysis in various fields including Psychological measurements can derive benefit from the use of TW statistic, especially when testing the significance of the largest or other eigenvalues are involved.

2.3 Difficulty value and Discriminating value:

Difficulty value of i-th item (p_i) is the proportion of individuals giving correct answer to the item and is calculated as $p_i = \frac{C_i}{n}$ where C_i denotes number of persons giving correct answer to the i-th item and *n* is the total number of persons taking the test. Clearly,$0 \le p_i \le 1$. To maximize test reliability, items with homogeneous item-difficulty could be preferred, but most standardized tests use items showing wide range of difficulty values. Discriminating value of i-th item $Disc_i$ is traditionally calculated as $\frac{UG-LG}{n}$ where UG denotes number of persons in the upper 27% who answered the item correctly and LG denotes the lower 27% who correctly answered the item. However, assessing quality of items or test ignoring 46% of data is not desirable.

$Disc_i$ indicates ability of an item to distinguish between examines with high trait level and with low trait level (Ferrando, 2012). Discriminating value of a test/scale reflects ability to differentiate or discriminate the sample from “preferred” and “non-preferred” or “high achievers” and “poor achievers”, etc. $Disc_i$ depends on whether the item is easy or difficult or moderately difficult. Relationship between item difficulty values $(Diff_i)$ of of MCQ type test based on the entire data and $Disc_i$ based on 54% of the data gave contrasting results. While Rao, et al. (2016) found $r_{Diff_i,Disc_i}$ = 0.56, Sim and Rasiah (2006) found positive $r_{Diff_i,Disc_i}$ at the “easy end” (percentage $Diff_i \geq$ 80%), and negative value at “difficult end” (percentage $Diff_i\ \leq$ 20%) and dome-shaped curve when all items are considered. Negative $Disc_i$ indicates that more number of individuals in LG could answer the i-th item correctly than the number of individuals in UG.

Item discrimination by point-biserial coefficient (r_{pbs}) shows degree to which an individual item is measuring the same thing as the rest of the items. Items with low or negative discrimination values may be rewarded or deleted. Classification of items as “good”, “fair” and “poor” based on discrimination value ≥ 0.30, between 0.10 and0.30; and ≤ 0.10 respectively appears to be arbitrary and may be questioned.

Non-availability of relationship between $Diff_i$ and $Disc_i$ and their relationships with test parameters fail to reflect impact of deletion of one or more items on test reliability (r_{tt}) or discriminating value of the test $(Disc_T)$ or difficulty value of the test $(Diff_T)$. Chauhan, et al. (2013) suggested further study to investigate correlation between difficulty index and discriminative index. For a MCQ test with m-items (1 for correct answer and 0 otherwise) administered to n-respondents, Chakrabartty (2021) considered an item score follows Binomial distribution with parameters n and p where p is the probability of correct answer and is equal to $Diff_i = \frac{k}{n}$ where k denotes number of correct answer to the i-th item and mean and SD of the item are np and $\sqrt{npq}$ respectively, where $q = 1 - p = \frac{n-k}{n}$ and proposed following measures:

- $Diff_i = \frac{k}{n}$ and $0 \leq Diff_i\ \leq 1$

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

- $Disc_i$ = Coefficient of variation (CV) = $\frac{SD}{Mean} = \frac{S_{X_i}}{\overline{X_\iota}} = \frac{\sqrt{npq}}{np} = \sqrt{\frac{q}{np}} = \sqrt{\frac{q}{k}} = \sqrt{\frac{n-k}{nk}}$

$${Disc_i}^2 = \frac{n}{k} - 1 = \frac{1}{Diff_i} - 1 = \frac{1-Diff_i}{Diff_i}$$

$Disc_T = \frac{S_X}{\bar{X}}$ = CV of test scores and $0 \leq Disc_T\ \leq 1$

The following may be noted:

1. 0 ≤ $Diffi$≤1. $Diff_i$ increases monotonically with increase in k. The $Diff_i$ curve is positively slopped.

2. $Disc_i = 0 \Longrightarrow n = k$ i.e. the i-th item was passed by all the subjects. $Disc_i = 1 \Longrightarrow k = \frac{n}{n+1}$ which is a fraction. Thus, $0 \leq Disc_i < 1$. $Disc_i$ decreases with increase in k. Thus, $Disc_i$ curve is negatively sloped

3. If $\overline{X_\iota} = \overline{X_J}$ ($i \neq j$), the item with lower SD will have lower CV and lower$Disc_i$.

4. Sum of item variances $\sum_{i=1}^{m} S_{X_i}^2 = \sum_{i=1}^{m} \overline{X_\iota}^2 Disc_i^2$ ⟹Cronbach $\alpha = \frac{m}{m-1}(1 - \frac{\sum_{i=1}^{m} \overline{X_\iota}^2 Disc_i^2}{\bar{X}^2 Disc_T^2})$

5. Intersection of $Diff_i$ curve and $Disc_i$ curve (k_0) is the solution of $Diff_i = Disc_i$ i.e. $\frac{k}{n} = \sqrt{\frac{n-k}{k}} \Leftrightarrow k^3 = n(n-k)$

k_0 could be taken as the value (to the nearest integer) where the negatively slopped $Disc_i$ curve intersects with the positively slopped $Diff_i$ curve. The point k_0 is the central point for getting acceptance region of items says ($k_0 \pm \Delta$) where Δ could be chosen as SD or 2SD of distribution of item difficulty values or item discriminating values, depending on original number of items, type of test, whether to measure single dimension or multi dimensions and also considering relationship between test discrimination and test reliability.

Empirically, Chakrabartty (2021) found correlation between $Diff_i$ and $Disc_i$ = (-) 0.58.

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

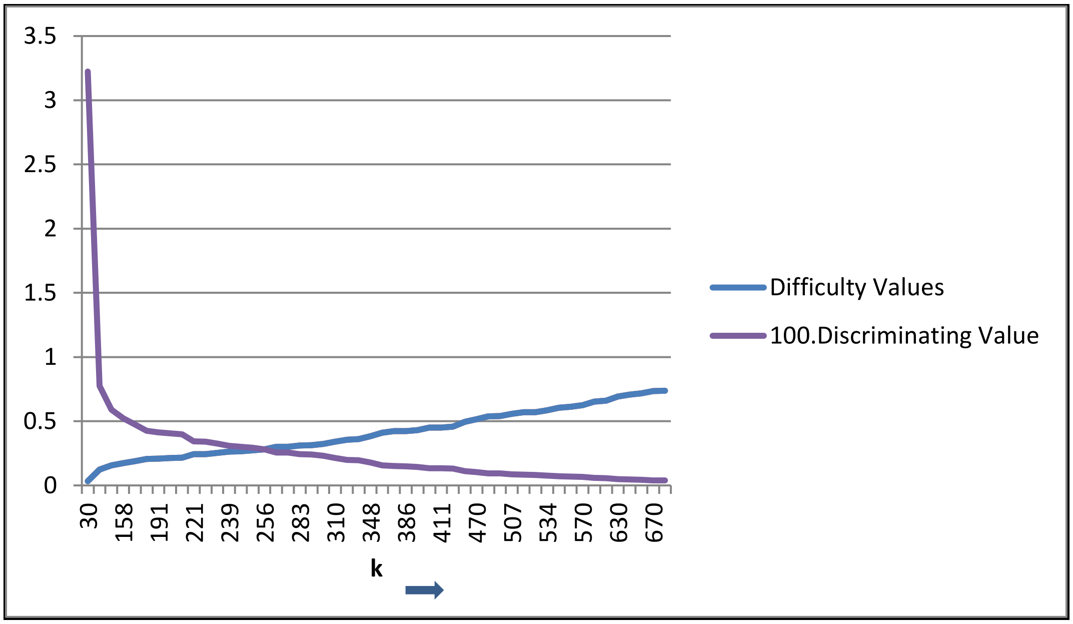

Figure 1. Item difficulty values and percentage discriminating values of items

Shifting k_0 to the right will increase proportion of items with high difficulty values (and low discriminating values). Thus, value of k_0 may be considered while deleting number of items from the test.

Chaktabartty (2020) extended the concept of $Disc_i$ and $Disc_T$ to scale containing Likert items from a single administration. The author compared seven dissimilarity measures Euclidian distance, Kullback- Leibler measure, Angular separation, Bhattacharyya measure, Hellinger discrimination, Chi-square measure and Coefficient of variation and found measures given by CV performs best in terms of satisfaction of desired properties. In addition, unbiased estimation of population CV is possible for normally distributed data (Sokal and Rohlf, 1995). Statistical inferences for CV for normally distributed data are often based on McKay‟s Chi-square approximation for the CV (Forkman, 2013).

3. Relationships

Mathematical relationships among item statistics and test parameters are derived as follows:

3.1 Difficulty and discriminating values:

i) $Diff_T$ and $Diff_i^{'}s$: $Diff_T = \frac{\sum_{i=1}^{m} Diff_i}{m}$ (11)

ii) $Diff_i$ and $Disc_i$: $Disc_i^{\,2} = \frac{n}{k} - 1 = \frac{1}{Diff_i} - 1 = \frac{1-Diff_i}{Diff_i}$ (12)

iii) $Diff_T \,.\, Disc_T = \frac{S_X}{m}$ (13)

iv) $r_{tt}\ (Disc_T)^2 = \frac{S_T{}^2}{S_X{}^2}\frac{S_X{}^2}{\bar{X}^2} = (\frac{S_T}{\bar{X}})^2 = (\frac{S_T}{\bar{T}})^2$ since $\bar{X} = \bar{T}$ (14)

Thus, product of test reliability and square of test discriminating value is equal to square of CV of true scores. $S_T{}^2$can be obtained as $S_X{}^2 - S_E{}^2$ where $S_E{}^2$ is given by equation (4) or by (6) for parallel tests.

Equation (14) shows that r_{tt} is inversely proportional to $Disc_T{}^2$. Thus, it is a non-linear negative relationship between reliability and discriminating value of a test. The equation does not involve number of items. However, addition or deletion of items may affect test variance (S_X^2) which in turn affect both r_{tt} and $Disc_T$. Appropriate selection of items can be made to have reasonable levels of $Disc_T$ and r_{tt}. For example, consider the figure 1.Here deletion of items with $100Disc_i \leq k_0$ will increase $Disc_T$ and may give a reasonable level of r_{tt}. For scale with m-number of k-point items, the items with low values of item variance can be deleted. Effect of deletion of items in stages and resulting effect on r_{tt} and $Disc_T$ and also on Cronbach α [expessed as $= \frac{m}{m-1}(1 - \frac{\sum_{i=1}^{m} \bar{X_i}^2 Disc_i^2}{\bar{X}^2 Disc_T^2})$] may be investigated empirically.

r_{pbs} reflects item-total correlation between an item score (dichotomous variable) and test score (continuous scale).Chakrabartty (2021) derived r_{pbs} for the *i*-th item in terms of item statistics as

$$r_{pbs(i)} = \frac{(M_{pi}-M_{qi})\sqrt{Diff_i(1-Diff_i)}}{\bar{X}\, Disc_T} \tag{15}$$

where M_{pi}: Test mean for persons answering the *i*-th item correctly

M_{qi}: Test mean for persons answering the *i*-th item incorrectly

Equation (15) shows negative relationship between item-total correlation ($r_{pbs(i)}$) and $Disc_T$. High value of $r_{pbs(i)}$ indicates that subjects who correctly answered the *i*-th item have done well overall on the test. Thus, $r_{pbs(i)}$ could be taken as measure item reliability. Clearly, $r_{pbs(i)} \geq 0$ if $\left(M_{pi} \geq M_{qi}\right)$. Items with negative or marginal value of $r_{pbs(i)}$ may be reworded or deleted.

3.2 Reliability and Factorial validity (FV):

$$\text{FV} = \frac{\lambda_1}{\sum\lambda_i} = \frac{\lambda_1}{Sum\ oftraces\ of\ the\ variance-covariance\ matrix} = \frac{\lambda_1}{Sum\ of\ item\ variances}$$

If item wise equidistant scores (E_i) are standardized to $Z_i \sim N(0,1)$, sum of eigenvalues = number of original variables (m).

For a test with m-number of standardized items:

- $FV_{Z-scores} = \frac{\lambda_1}{m}$
- Test variance (S_X^2) = m + $2\sum_{i\neq j=1}^{m} Cov\left(X_i, X_j\right)$
- $r_{tt} = \frac{S_T^2}{S_X^2} = \frac{S_T^2}{\sum\lambda_i + 2\sum_{i\neq j=1}^{m} Cov(X_i,X_j)} = \frac{S_T^2}{\frac{\lambda_1}{FV} + 2\sum_{i\neq j=1}^{m} Cov(X_i,X_j)}$ (16)

Equation (16) gives non-linear relationship between theoretical reliability and factorial validity of standardized scores, where each term of the denominator can be estimated from data and S_T^2 may be computed as S_X^2- S_E^2 where S_E^2 is given by (4).

Relationship between factorial validity and reliability as per α_{PCA} can be obtained considering equation (2)

$$\alpha_{PCA} = \left(\frac{m}{m-1}\right)\left(1 - \frac{1}{\lambda_1}\right) = \left(\frac{m}{m-1}\right)\left(1 - \frac{1}{FV.\sum\lambda_i}\right) = \left(\frac{m}{m-1}\right)\left(1 - \frac{1}{m.FV_{Z-scores}}\right) \tag{17}$$

(17) indicates higher value of $FV_{Z-scores}$ increases α_{PCA}

4. Other measures

4.1 Responsiveness:

Responsiveness deals with change scores ($\nabla X = X_t - X_{(t-1)}$) of the instrument of interest (X). Illustrative hypotheses (situations) to assess responsiveness are:

1. Correlations between ∇X and ∇Y of another instrument (Y) measuring similar constructs (strong relationships, say$>$ 0.5) or ∇Z of instrument measuring unrelated constructs (weaker relationships, say < 0.3) (Prinsen, et al. 2018)
2. Expected differences in ∇X between different subgroups (known groups) like say persons receiving an intervention of known efficacy and persons waiting for the intervention.
3. Magnitude of ∇X expected after undergoing a treatment with known efficacy on the construct of interest (medium effect size between 0.3 and 0.5)

However, due to different contents and many arbitrary hypotheses, responsiveness is never perfect (Mokkink et al. 2021). In addition, it is necessary to know distribution and behavior of ∇X.

Responsiveness of an instrument can be assessed using normally distributed P-scores of a test/scale. Percentage change of the j-th individual in the t-th period over $(t-1)$-th period is given by

$$\frac{P_{jt}-P_{j(t-1)}}{P_{j(t-1)}} \times 100 \tag{18}$$

Similarly, for a sample, responsiveness is reflected by

$$\frac{\overline{P_t}-\overline{P_{(t-1)}}}{\overline{P_{(t-1)}}} \times 100 \tag{19}$$

While (18) focuses on individual level, (19) deals with sample/group level. Positive value of (18) implies progress of the j-th individual in successive time periods, assuming positive relation of each item score with test score. Similarly, $\frac{\overline{P_t}-\overline{P_{(t-1)}}}{\overline{P_{(t-1)}}} \times 100 > 0$ quantifies progress made by the group of individuals in the t-th period over the previous year. Negative value of (18) or (19) indicates deterioration and calls for attentions to initiate necessary corrective action. Based on (18) individuals can be ranked with respect to extent of progress registered by them. It is possible to test significance of progress or deterioration since ratio of two normally distributed variable follows χ^2 distribution.

Progress path of one or a sample of individuals across time i.e. trajectory over time can be computed considering (18) and/or (19) at various values of t =0(base period or starting period), 1, 2, 3, ……..and so on. Such progress path is analogous to hazard function and can be used to compare progress pattern of individuals or groups showing effectiveness in teaching-learning environment or treatments/cares from the start for continuous and comprehensive evaluation. Such trajectories can help to identify high-risk groups and contribute to monitor progress on education.

4.2 Integrating scales:

Equivalent scores are required to see whether qualifying marks or cut-off marks of two or more tests/scales are same in classifying individuals as “Passed” or Failed” or “with disease” and “without diseases” where tests or scales have different test lengths, scores ranges and score distributions. Regression equation of the form $Y = \alpha_1 + \beta_1 X$ is different from the regression of X on Y i.e. $X = \alpha_2 + \beta_2 Y$. Thus, relationship between X and Y will not be unique. Moreover, predicting and equating are different concepts.

P-scores $\sim N(\mu, \sigma)$help to integrate two scales X and Y with pdf $f(x)$ and $g(y)$ in terms of equivalent scores $(x_{0,}\, y_0)$ given by $\int_{-\infty}^{x_0} f(x)dx = \int_{-\infty}^{y_0} g(y)dy$ for a given value of say x_0 i.e. area of the curve $f(x)$ up to x_0= area of the curve $g(y)$ up to y_0 which can be solved by using standard Normal table, even if the scales differ with respect to number of items or dimensions (Chakrabartty, 2021b), who found correlation between such equivalent scores exceeded 0.99. This avoids the problems of linear equating or percentile equating.

5. Discussions

The paper describes methods of finding difficulty value $(Diff_T)$, discriminating value of test or scale $(Disc_T)$ and item $(Disc_i)$ as Coefficient of variation (CV), considering the entire data and derives relationships among them including non-linear negative relationships between theoretically defend test reliability $(r_{tt-theoretical})$ and $Disc_T$. Thus, one cannot increase both reliability and discriminating value of a test. Cronbach alpha was expressed as function of item difficulty values and test discriminating value.

Values of $r_{tt-theoretical}$, variance of true score (S_T^2) and error score (S_E^2) can be found by dichotomization of the test in two parallel sub-tests and considering lengths of vectors representing the sub-tests$\|X_g\|$, $\|X_h\|$ and angle between the vectors $\boldsymbol{X_g}$ and $\boldsymbol{X_h}$. $r_{tt-theoretical}$ can be used to find reliability of a battery of tests after defining battery scores appropriately.

Item reliability of MCQ type test in terms of point bi-serial correlation (r_{pbs}) runs the risk of being negative if mean for persons answering the *i*-th item correctly < mean for persons answering the *i*-th item incorrectly. Inter-item reliability as average of inter-item correlations suffers from limitations since correlations are not additive. Methods suggested to make correlations additive or to compute mean and SD of several correlations.

Relationships of factorial validity (FV) = $\frac{First\ eigen\ value}{Sum\ of\ eigen\ values}$ with $r_{tt-theoretical}$ and also with α_{PCA} derived. The former gives non-linear relationship between theoretical reliability and factorial validity of standardized scores $(FV_{Z-scores})$, where each term can be estimated from data. However, higher value of $FV_{Z-scores}$ increasesα_{PCA}. Significance of FV can be found since FV = $\frac{\lambda_1}{\sum \lambda_i}$ follows a Tracy–Widom (TW) distribution.

Normally distributed P-scores enabling better arithmetic aggregation, meaningful comparisons and satisfying desired properties contribute to improve scoring of instruments including parametric analysis like PCA, testing and estimation of mean, variance of item, test or scale, coefficient of variation (CV), Cronbach alpha at population level, factorial validity in terms of eigen values, Reliability by α_{PCA}, testing of $H_0: r_{tt} = 1$, finding equivalent scores of two or more instruments.

Such scores also help to quantify responsiveness of an instrument at individual level or at sample level where positive value indicates progress. In the former case, individuals can be ranked with respect to extent of progress registered by them. It is possible to test significance of progress or deterioration since ratio of two normally distributed variable follows χ^2 distribution. Progress path of one or a sample of individuals across time can be found by the proposed measure of responsiveness. Such progress paths reflect effectiveness in teaching-learning environment or corrective action from the start and also to identify high-risk groups and contribute to monitor progress in education.

6. Conclusions:

Analysis of data emerging from MCQ type tests or scales/questionnaire containing items with different number of response-categories may be transformed to follow normal distribution with estimated parameters. Derived relationships of test reliability with discriminating value or with factorial validity show that all the test parameters cannot be improved simultaneously. However, such relationships have potentials to find optimal value of one parameter to maximize another parameter. Future studies may explore such potentials with empirical investigations.

Declaration:

Credit statement: Conceptualization; Methodology; Analysis; Writing and editing the paper by the Sole Author

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

Conflicts of interest: The author has no conflicts of interest to declare.

Availability of data and material: Nil (The paper used hypothetical data).

Code availability: No application of software package or custom code.

References

ACT (1997): ACT assessment: Technical manual. Iowa City, Iowa.

Berchtold A (2016): Test–retest: Agreement or reliability? Methodological Innovations 9: 1–7.

Bewick, Viv; Cheek, Liz and Ball, Jonathan (2003): Statistics review 7: Correlation and regression, Critical Care; 7:451-459. DOI 10.1186/cc2401

Chakrabartty, S.N. (2022): Disability and Quality of Life. Health Science Journal, 16(12); 1 - 6 DOI: 10.36648/1791- 809X.16.12.989

Chakrabartty, S.N. (2021): Assessment of Item and Test parameters: Cosine Similarity Approach. International Journal of Psychology and Education Studies; 8 (3), 28-38, https://dx.doi.org/10.52380/ijpes.2021.8.3.190

Chakrabartty, Satyendra Nath (2021b): Integration of various scales for Measurement of Insomnia. Research Methods in Medicine & Health Sciences, 2(3), 102-111, 10.1177/26320843211010044

Chakrabartty, Satyendra Nath (2020): Discriminating value of Item and Test. International Journal of Applied Mathematics and Statistics; 59(3), 61 - 78

Chauhan, PR, Ratrhod, SP, Chauhan, BR, Chauhan, GR, Adhvaryu, A. and Chauhan, AP (2013): Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot, BIOMIRROR, Vol.4 (06), 1-4.

Cohen R, Swerdlik M.(2010): Psychological testing and assessment. Boston: McGraw-Hill Higher Education.

Cortina, J. (1993): What is coefficient alpha? An examination of theory and methods. Journal of Applied Psychology 78:1, 98-104.

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

De Hooge IE, Zeelenberg M., Breugelmans SM (2007): Moral sentiments and cooperation: Differential influences of shame and guilt. Cognition and Emotion; 21:1025–1042. doi: 10.1080/02699930600980874.

Ebel, RL and Frisbie, DA (1991): Essentials of educational measurement. Prentice Hall of India Pvt. Ltd.

Edmonds, GW, Goldberg, LR, Hampson, SE & Barckley, M (2013): Personality stability from childhood to midlife: Relating teachers’ assessments in elementary school to observer and self-ratings 40 years later. Journal of Research in Personality, 47, 505–513. doi:10.1016/j.jrp.2013.05.003

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

Field, A. P. (2003): The problems in using fixed-effects models of meta-analysis on real-world data. Understanding Statistics, 2, 105–124.

Forkman, J. (2013): Estimator and tests for common co-efficient of variation in Normal Distribution. Communications in Statistics – Theory & Methods, 38(2), 233-251

Garcia, Edel (2012): The Self-Weighting Model, Communications in Statistics - Theory and Methods, 41:8, 1421-1427. http://dx.doi.org/10.1080/03610926.2011.654037

Hampson, SE, Edmonds, GW, Barckley, M, Lewis, R, Goldberg, Joan P. Dubanoski & Teresa A. Hillier (2015): A Big Five approach to self-regulation: personality traits and health trajectories in the Hawaii longitudinal study of personality and health, Psychology, Health & Medicine, DOI: 10.1080/13548506.2015.1061676

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.130-159). American Council on Education.

Huang, Rui-Ting and Tang, Tzy-Wen (2013): Examining the role of gender differences in mobile English Learning. International Journal of Instructional Technology and Distance Learning; 10(8), 43 - 51

Jelenchick LA, Becker T, Moreno MA (2012): Assessing the psychometric properties of the Internet Addiction Test (IAT) in US college students. Psychiatry Research 196: 296–301

Kehoe, Jerard (1994): Basic Item Analysis for Multiple-Choice Tests, Practical Assessment, Research, and Evaluation; Vol. 4, Article 10. DOI: https://doi.org/10.7275/07zg-h235

Kristof, W. (1963): The statistical theory of speeded-up reliability coefficients when a test has been divided into several equivalent parts. Psychometrika, 28, 221–238.

Lantz, B. (2013): Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations. Electronic Journal of Business Research Methods; 11, 16-28

Larsen, RJ, & Marx, ML (2010): Introduction to Mathematical Statistics and Its Applications (5 edition.). Boston: Pearson.

McDonald, RP. (1999): Test theory: A unified treatment. Mahwah (NJ); Lawrence Earlbaum Associates, Inc.

Mokkink, Lidwine; Terwee, Caroline and de Vet, Henrica (2021): Key concepts in clinical epidemiology: Responsiveness, the longitudinal aspect of validity, Journal of Clinical Epidemiology, Vol.140, 159-162, https://doi.org/10.1016/j.jclinepi.2021.06.002.

Nadler, Boaz (2011): On the distribution of the ratio of the largest eigenvalue to the trace of a Wishart matrix. Journal of Multivariate Analysis, 102; 363-371

Parkerson, HA. Noel, M., Gabrielle MP, Fuss, Samantha, Katz, J., Gordon JG. Asmundson (2013): Factorial Validity of the English-Language Version of the Pain Catastrophizing Scale–Child Version, The Journal of Pain, 14 (11), 1383-1389, https://doi.org/10.1016/j.jpain.2013.06.004

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

Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwee CB. (2018): COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 27(5):1147-1157. doi: 10.1007/s11136-018-1798-3.

Pronk T, Molenaar D, Wiers RW, Murre J. (2022): Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. Psychon Bull Rev. 29(1):44-54. doi: 10.3758/s13423-021-01948-3.

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: 201-204

Rao CR (1973): Linear Statistical Inference and Its Application. 2nd ed. New Delhi: Wiley Eastern Private Limited.

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

Sim, Si–Mui and Rasiah, RI.(2006): Relationship Between Item Difficulty and Discrimination Indices in True/FalseType Multiple Choice Questions of a Para-clinical Multidisciplinary Paper, Annals of the Academy of Medicine, Singapore, 35(2), 67-71.

Sokal, RR and Rohlf, FJ (1995): Biometry: The Principles and Practice of Statistics in Biological Research (3rd Ed.), New York: Freeman

Ten Berge, JMF & Hofstee, WK (1999): Coefficients alpha and reliabilities of unrotated and rotated components. Psychometrika, 64, 83–90. doi:10.1007/BF02294321

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.

Ogasawara, H. (2006): Approximations to the distribution of the sample coefficient alpha under non-normality. Behaviormetrika; 33(1), 3–26

Utwin MS (1995): How to Measure Survey Reliability and Validity. Sage Publications, Thousand Oaks

Young KS (1998): Caught in the Net: How to Recognize the Signs of Internet Addiction—And a Winning Strategy for Recovery. New York: John Wiley & Sons

Zimmerman, DW, Zumbo, BD. and Williams, RH (2003). Bias in estimation and hypothesis testing of correlation. Psicológica 24:133–158.