**Edukasiana: Jurnal Inovasi Pendidikan**
ISSN 2809-6770
Volume 2, Issue 4, 2023, pp. 285-293

# Item Deletions Based on Difficulty Values and Discriminating Values

**Satyendra Nath Chakrabartty***
Indian Statistical Institute, India
*Corresponding Author: chakrabarttysatyendra3139@gmail.com

**Article History:**
Received 2023-07-11
Accepted 2023-09-20

**DOI:**
10.56916/ejip.v2i4.455

***ABSTRACT***

***Background:*** *Deletion of items from MCQ tests or Likert type scales may be necessary due to various reasons.* ***Methods:*** *Considering entire data the paper gives new measures of difficulty and discriminating value of items as well as test along with their relationships including relationship with test reliability ($r_{tt}$). Discriminating value of test ($Disc_T$) and item ($Disc_i$) are expressed as coefficient of variation (CV) of test scores and item scores respectively.* ***Results:*** *Non-linear relationship between $Disc_i$ and $Diff_i$ derived. As number of persons giving correct answer to an item increases, $Diff_i$ curve increases and $Disc_i$ curve decreases and intersect at a point (say $k_0$). Items lying outside the interval $[k_0 \pm 2SD]$ where SD is standard deviation of $Diff_i$ or $Disc_i$ can be deleted. Choosing acceptance region as $[k_0 \pm 3SD]$ may result in discarding too few items. For Likert scale, items with high values of CV may be deleted. Relationship of reliability and discriminating values helps to find effect of such deletions.* ***Conclusions:*** *Proposed method of item deletions based on difficulty values and discriminating values offers significant benefits and is recommended. However, the approach may be compared with deletion of items by "alpha if the item is deleted". Future studies suggested.*

***Keywords:*** *item deletion; difficulty value; discriminating value; coefficient of variation; reliability and validity*

## 1. INTRODUCTION

Types of tools used in Educational research are Multiple Choice Questions (MCQ) tests scored as "1" for correct answer to an item and "0" for the rest items; Likert type scales to monitor student learning for feedback and also to assess the important outcomes at the end of the instructions. Each such tool uses summative scores obtained as sum of item scores.

It may be necessary to delete some of the items due to various reasons. Improvement of test requires deleting ineffective items or items with only few corrected answers i.e. extremely difficult items. The existing test may be lengthy or deletion of items may increase reliability of the test. Similarly, deletions of number of items in a questionnaire are important to have reduced response error, higher respondent engagement, reduction of multicollinear items improved test charecteristics.

Traditional approach is to consider item-analysis results and delete or modify items based on item difficulty value and item discriminating value. Difficulty value of an item ($Diff_i$) is defined as the proportion of correct responses to the item and discriminating value of an item ($Disc_i$) indicates ability of the 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 based on top 27% and bottom 27% of data which amounts to rejection of 46% of the data and hence not desirable. For the $i$-th item, relationship between $Diff_i$ based on the entire data and $Disc_i$ based on 54% of the data is difficult

to interpret and may give rise to contrasting results. For example, Rao, et al. (2016) found $r_{Diff_i,Disc_i}$ = 0.56 which contradicts usual idea of poor discrimination value of a very easy items (high difficulty value) which was answered correctly by most of the subjects taking the test. Sim and Rasiah (2006) found positive value of $r_{Diff_i,Disc_i}$ for $Diff_i$ ranging between 0.80 to 1.00 and negative correlations when $0 \leq Diff_i \leq 0.20$ and relationship showed dome-shape when all the items are considered. Further study to investigate correlation between $Diff_i$ and $Disc_i$ was proposed (Chauhan, et al. 2013). Clearly, better evaluations of effectiveness of MCQ items are needed. Absence of clear 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 parameters like reliability ($r_{tt}$), item-total correlation by point bi-serial correlation ($r_{pbs}$), discriminating value of the test ($Disc_T$) or difficulty value of the test ($Diff_T$).

The paper considers entire data and quantifies difficulty and discriminating value of items and tests and their relationships including relationship with test reliability, as per definition (ratio of true score variance and observed score variance) from a single administration.

**Literature survey:**

Deletions of items are usually done by following one or more approaches given below:

1. Low value of discriminating index computed as difference between the top $\frac{1}{3}$rd of respondents and the bottom $\frac{1}{3}$rd of respondents.
2. Low correlations between an item and the total score.
3. Items whose deletion improves Cronbach's alpha i.e. alpha if item is deleted
4. Items with low factor loadings

**Problem areas and issues:**

*Approach 1*: The approach suffers from disadvantages of not considering entire data and giving rise to contrasting results.

*Approach 2*: Researchers differed in deciding such value of correlation. While Avanoor and Mahendran, (2018) suggested to delete an item if the correlation is less than 0.3, Kehoe (1995) and Popham (2011) favoured deletion of an item if $r_{pbs} \leq 0.15$ and item-total correlation is less than 0.19.

*Approach 3:* To find "alpha if the item is deleted" and delete the items accordingly so that the test excluding the deleted items has higher value of alpha. In other words, delete the *j*-th item if $\alpha_{j-1} > \alpha_j$ where $\alpha_j$ denotes reliability in terms of Cronbach alpha of the test including the *j*-th item and $\alpha_{j-1}$ denotes reliability of the test without the *j*-th item. If deletion of an item increases alpha for the test, the item needs to be deleted (Raykov, 2008). However, such modified test may lower criterion validity (Raykov, 2007). In addition, set of items showing high value of alpha may not always be homogenous or uni-dimensional (Green et al. 1977).

Test reliability does not indicate the degree of discrimination offered by an instrument (Hankins, 2007). If items with $Disc_i \leq 0$ are included, measurement disturbance by the test may occur. Thus, $Disc_i$ or $Disc_T$ are closely related to the quality of the score as a measure of the trait (McDonald, 1999). Range of item discrimination index is between - 1.0 to 1.0. (Shakil, 2008; Denga, 1987) and is not defined if all subjects taking the test got same score on the item.

Erhart et al. (2010) investigated item deletion to maximize alpha and item fit of the partial credit model (Masters, 1982) and opined that item deletion approaches need to consider additional analyses since quality of a test is more than test reliability.

Major issues on test reliability in terms of Cronbach $\alpha$ and validity as correlation between test scores and scores of a chosen criterion scale are as follows:

1) Alpha as a measure of internal consistency is concerned with the homogeneity of the items within a test and does not work well for a multi-dimensional test.
2) Alpha assumes uncorrelated errors and **tau-equivalent items which imply** all the **factor loadings** are **same** (Ogasawara, 2006). However, equality of **factor loadings** is rather rare for tests used in educational research (Pronk et al. 2022).
3) If items are not essentially tau-equivalent and the test measure different constructs i.e. multi-dimensional test, alpha may get distorted. However, many scales reports alpha despite finding several factors from PCA or FA.
4) Huang et al. (2021) found that the construct with highest eigenvalue had the maximum alpha. Using results of PCA, Ten Berge and Hofstee, (1999) proposed test reliability as $\alpha_{PCA} = (\frac{m}{m-1})\,(1-\frac{1}{\lambda_1})$ where $\lambda_1$ is the first (largest) eigenvalue of correlation matrix of *m*-number of items.
5) Clearly, different methods of finding reliability deviating from definition of reliability may give different values of reliability even from the same data. Chakrabartty (2021) proposed finding theoretical reliability ($r_{tt-Theoretical}$) as per its definition from single administration of a test with *m*-items as

$$r_{tt-Theoretical} = 1 - \frac{\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\|X_h\|Cos\theta_{gh}}{n{S_X}^2} \quad (1)$$

where the test is dichotomized to two parallel sub-tests (*g*-th and *h*-th) each with $\frac{m}{2}$ items, $\|X_g\|$ and $\|X_h\|$ are length of the sub-test vectors, computed as $\|X_g\| = \sqrt{\sum_{i=1}^{m/2} {X_{ig}}^2}$ and $\|X_h\| = \sqrt{\sum_{i=1}^{m/2} {X_{ih}}^2}$ and $\theta_{gh}$ is the angle between the $\boldsymbol{X_g}$ and $\boldsymbol{X_h}$.
6) Different selections of criterion scale may give different values of validity of a test/scale.
7) Construct validity is difficult to interpret when a test is multi-dimensional. Question may arise it is validity for which factor?
8) Parkerson et al. (2013) suggested to find Factorial validity ($V_{Factorial}$) for the main factor for which the scale was developed as $V_{Factorial} = \frac{\lambda_1}{\sum \lambda_i}$ where $\lambda_1$ denotes the highest eigenvalue corresponding to the main factor for which the scale was developed. $\sum \lambda_i$ is the sum of eigenvalues = trace of the variance-covariance matrix = Sum of item variances. Clearly, $V_{Factorial}$ will be high for uni-dimensional tests.
9) It is possible to find relationship between $\alpha_{PCA}$ and $V_{Factorial}$ since both are functions of $\lambda_1$ is the first (largest) eigenvalue.

*Approach 4:* Muñiz & Fonseca-Pedrero (2019) opined that Exploratory Factor Analysis (EFA) is most appropriate model for item analysis, where inappropriate items are discarded. Major assumptions of EFA include linearity (nonlinear relationships may not be accurately captured by factor analysis); multivariate normality (significant departures from normality may affect the accuracy of the results); absence of outliers, adequate sample size, etc. Violation of the assumptions like nonlinear relationships may not be accurately captured by factor analysis. Significant departures from normality may affect the accuracy of the results.

Measure of sampling adequacy (MSA) is useful for debugging inappropriate items before factor analysis (FA) is undertaken (Lorenzo-Seva and Ferrando, 2021) where MSA of *i*-th item of a test with *m*-items is computed as $MSA_i = \frac{\sum_{i\neq k}^{m} r_{ik}^2}{\sum_{i\neq k}^{m} r_{ik}^2 + \sum_{i\neq k}^{m} P_{ik}^2}$ where $r_{ik}$ is the correlation between i-th and k-th items and $p_{ik}$ is the corresponding partial correlation. Clearly, $0 \leq MSA_i \leq 1$. Value of $MSA_i$ closed to 1

indicates appropriateness of the item for FA. Low value of $MSA_i$ (closed to zero) may occur if the item does not belong to the same family as the other items or do not sample the same domains measured by the remaining items. Thus, $MSA_i \approx 0$ could imply either the $i$-th item is noisy lacking discriminating power (Ferrando, 2012) or the item is redundant and does not share the contrast being measured by the other items of the test. Cut-off value for discarding items may be relevant for noisy items but not for the redundant items. Presences of both noisy and redundant items create problems for EFA in item analysis. In addition, noisy items with poor loadings on any factor for multi-dimensional test fail to test whether the items measure different factors or are pure noise. However, deletion of an item will change mean, SD of the test/scale and also correlation of a retained item with total test/scale score.

## 2. DIFFICULTY AND DISCRIMINATING VALUES

### MCQ tests

Suppose a MCQ-test with $m$-items has been administered among $n$-subjects. Scores of the subjects can be presented as a $n$-dimensional vector $\boldsymbol{X} = (X_1, X_2, \dots \dots . X_n)^T$ where the component $X_i$ denotes test score of the $i$-th subject. Consider another $n$-dimensional vector $\boldsymbol{I}$ representing maximum possible score where each component is equal to 1. Let the angle between the vectors $\boldsymbol{X}$ and $\boldsymbol{I}$ be $\theta_X$. Here, $Cos_{\theta_X} = \frac{\sum_{i=1}^{n} X_i I_i}{\|X\|\|I\|}$ by definition boils down to

$$Cos_{\theta_X} = \frac{\sum_{i=1}^{n} X_i}{\|X\|\sqrt{n}} \text{ since } I_i = I_j \forall\, i, j = 1,2, \dots \dots, n \quad (2)$$

From (2), test men $\bar{X} = \frac{\|X\| Cos_{\theta_X}}{\sqrt{n}}$ (3)

and test variance $S_X^2 = \frac{\|X\|^2 Sin_{\theta_X}^2}{n}$ (4)

Considering the entire data, Chakrabartty (2021) proposed:

- Difficulty value of a test as $Diff_T = \frac{\|X\| Cos_{\theta_X}}{\|I\|} = \frac{\bar{X}}{m}$ (5)
- Difficulty value of an item as $Diff_i = Cos_{\theta_i}{}^2 = \frac{k}{n}$ (6)

where $k$ denotes number of persons answering the item correctly.

Clearly, $\bar{X} = \sum_{i=1}^{m} Diff_i$ and $Diff_T = \frac{\sum_{i=1}^{m} Diff_i}{m}$; $0 \le Diff_i \le 1$ and also $0 \le Diff_T \le 1$. Higher value of $Diff_i$ means the item is easy. Similarly, higher value of $Diff_T$ implies the test is easy.

- Discriminating value of a test $Disc_T = \frac{S_X}{\bar{X}} = CV_T$ (7)
  where $CV_T$ denotes Coefficient of variation of test
- Discriminating value of an item $Disc_i = \frac{S_{X_i}}{\bar{X}_i} = \sqrt{\frac{n-k}{nk}} = CV_i$ (8)
  where $CV_i$ denotes Coefficient of variation of i-th item

Here, $0 \le Disc_i < 1$ and not between [-1 to 1] as obtained from top 27% and bottom 27% of data. $Disc_i$ is maximum if $k = 1$ and minimum if $k = (n-1)$.

If two different items have same difficulty value, the item with lower SD will have lower CV and lower $Disc_i$ .

Chakrabartty (2021) derived the following relationships:

$$Disc_i{}^2 = \frac{1 - Diff_i}{n.Diff_i} = \frac{1 - Diff_i}{k} \quad (9)$$

$$r_{tt}.Disc_T{}^2 = (\frac{S_T}{\bar{X}})^2 \quad (10)$$

where $r_{tt} = \frac{S_T^2}{S_X^2}$

$$Diff_T.Disc_T = \frac{S_X}{m} \quad (11)$$

Equation (9) depicts relationship between $Disc_i$ and $Diff_i$ which is non-linear. For lower *k*, $Diff_i$ is reduced which tends to increases $Disc_i$. As per equation (10), test reliability ($r_{tt}$) and test discriminating values are negatively related in a non-linear fashion. Thus, it is not possible to increase both $r_{tt}$ and $Disc_T$ simultaneously.

Correlation between a binary item and total scale score is best measured by Point bi-serial correlation. For the $i$-th item, Chakrabartty (2021) derived

$$r_{pbs(i)} = \frac{(M_{pi}-M_{qi})\sqrt{Diff_i-(1-Diff_i)}}{\bar{X}Disc_T} \quad (12)$$

where $M_{pi}$ denotes the test is mean for persons answering the item correctly and $M_{qi}$ is the test mean for persons answering the $i$-th item incorrectly. Clearly, the relationship between $r_{pbs(i)}$ and $Disc_T$ is negative. If $r_{pbs(i)}$ is high, it means persons passing the $i$-th item have done well on the test.

**Deletion of items**

If *k*= 0 for an item, the item is extremely difficult and each subject fails to 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 forthwith. If $Disc_i = Disc_j$ for $i \neq j$ then item with higher SD is preferred to be retained.

Equation (9) depicts a non-linear relationship between item difficulty value and item discriminating value. Lower $Diff_i$ i.e. low value of $k \Longrightarrow$ higher$Disc_i$. Similarly, higher $Diff_i \Longrightarrow$ lower$Disc_i$. Thus, correlation between $Diff_i$ and $Disc_i$ will be negative. In other words, as $k$ increases, $Diff_i$ curve (or percentage$Diff_i$) will be positively sloped and $Disc_i$ curve (or percentage$Disc_i$) will be negatively slopped and the two curves will intersect at point ($k_0$)where $Diff_i$ = $Disc_i$. Value of $k_0$ can be obtained using equation (6) and (8) and by solving $\sqrt{\frac{n-k}{nk}} = \frac{k}{n}$ or $k^3 = n(n-k)$. Value of $k_0$ to be taken to the nearest integer.

Items may be retained by choosing the acceptance region as $[k_0 \pm 2SD]$ where SD is standard deviation of $Diff_i s$ or $Disc_i s$. Choosing acceptance region as $[k_0 \pm 3SD]$ may result in discarding too few items. In addition, considering skew of distribution of $Diff_i$(or $Disc_i$), few more items having high concentration at the tail may be discarded. It may be noted that deletion of one or more items will change values of $Diff_T$ & $Disc_T$.

Other considerations for item deletions are low value of point biserial correlation and alpha if item is deleted.

However, choice of acceptance region (or deletion region) 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 (equation 10). Discarding 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 product of $Diff_T$ & $Disc_T$ which is equal to SD per item. Effect of item deletions need to be checked with increase in test reliability and/or factorial validity.

**Likert scales:**

Concept of discriminating values of items and test in terms of coefficient of variation (CV) can be extended for Likert scales also where difficulty value is not relevant. Mean of a polytomous items is simply the average score. Chakrabartty (2020) compared seven dissimilarity measures which can be computed from a single administration of a questionnaire using proportion for each cell of the Item-Response category matrix and found that CV has maximum advantages to find discriminating values of Likert items and also for the Likert questionnaire. Here, $Disc_i = \frac{SD_i}{Mean_i}$ and $Disc_T = \frac{SD_{Test}}{Mean_{Test}}$. Lower value of CV is desirable. It is possible to estimate population CV and test statistical hypothesis on equality of $CVs$.

For a scale with $m$-items, relationship of Cronbach α and $Disc_T$ was derived as

$$\alpha = (\frac{m}{m-1})(1 - \frac{\sum_{i=1}^{m} \bar{X}_i{}^2.Disc_i{}^2}{\bar{X}^2.Disc_T^2}) \quad (13)$$

and $Disc_T^2 = \frac{CV_{True\ scores}{}^2}{r_{tt}}$ (14)

Each of (13) and (14) indicates negative relationship between test reliability and $Disc_T$ i.e. higher the $Disc_T$, lower is the reliability and vice versa.

Deletions of items of a Likert type test may be done by removing items with high values of CV (i.e. high value of SD). Reliability of the scale (or test) containing the retained items is likely to get improved because of negative relationship of reliability and discriminating values. For the same reason, items with low CV may be retained.

**Distribution and statistical tests of CV**

Scores of an item of MCQ test can be taken to follow Binomial distribution with parameters $n$ and $p_i$where $n$ denotes number of individuals taking the test and $p_i$ is the probability of success in a single trial i.e. difficulty value of the $i$-th item ($Diff_i$). Convolution of distributions of item scores will give the distribution of sum of all item scores which will also be Binomial.

$Disc_i$ is given by $CV_i$ which is equal to $\frac{S_{X_i}}{\bar{X}_i}$. CV can be used to compare discriminating value of two items even if they differ significantly with respect to mean. Similarly, discriminating value of two tests or scales can be compared on the basis of CV. Unbiased estimate of population CV for normally distributed data is $\widehat{CV} = (1 + \frac{1}{4n})(Sample\ CV)$ (Sokal and Rohlf, 1995). Asymptotic test for equality of CV proposed by Feltz and Miller (1996) consider test statistic which follows $\chi^2$distribution and is widely used. However, the 'Modified signed-likelihood ratio test (SLRT) for equality of CVs for different sample sizes (Krishnamoorthy and Lee, 2013) has more advantages. Software package for testing equality of CVs from multiple groups is given by Marwick and Krishnamoorthy, (2019).

For meaningful comparisons of tests with different response-categories and undertaking estimation and statistical testing, $Disc_i, Disc_T$, reliability and validity may be computed after transforming each ordinal item scores to equidistant, normally distributed scores in the same sore range say [1,100] by the method proposed by Chakrabartty (2022).

## 3. DISCUSSION

Under classical test theory (CTT), new measures of discriminating value of item ($Disc_i$) and test ($Disc_T$) and also difficulty value of test ($Diff_T$) are given considering entire data. Such newly defined measures and their relationships were derived including relationship with test reliability, as per definition. All the measures and relationships can be computed from a single administration of the test or scale.

$Diff_i$ of MCQ test with $n$-number of items is in line with usual notion of difficulty value which actually measures degree of easiness of a test. Here, $0 \leq Disc_i < 1$ and $0 \leq Disc_T < 1$. Range of $Diff_i$ is from 0 to 1. As number of correct answer to items ($k$) increases, positively sloped percentage$Diff_i$ curve will and negatively slopped percentage$Disc_i$ curve will intersect at point ($k_0$)where $Diff_i = Disc_i$ i.e. $k_0$ is the solution of the equation $\sqrt{\frac{n-k}{nk}} = \frac{k}{n}$ or $k^3 = n(n-k)$. Items lying outside $[k_0 \pm 2SD]$ may be deleted where SD is standard deviation of $Diff_i s$ or $Disc_i s$.

$Disc_i$ and $Disc_T$ for Likert type tests in terms of SD per mean (i.e. CV) has desired properties. Items with high values of CV may be deleted. In addition, few more items having high concentration at the tail may be discarded. It may be noted that deletion of one or more items will change values of $Diff_T$ & $Disc_T$.

Effect of deletion of items needs to be investigated using the derived relationships among the proposed measures with emphasis on test reliability as per theoretical definition which is negatively related to $Disc_T$ in non-linear fashion. Effect of deletion of items on validity or factorial validity may be investigated by undertaking PCA with the retained items. Methods of estimation of population CV and statistical testing of hypothesis on equality of two or more CVs are also suggested.

## 4. CONCLUSION

The proposed method of item deletions based on difficulty values and discriminating values offers significant benefits and is recommended. However, the approach may be compared empirically with deletion of items by "alpha if the item is deleted" with respect to an optimal range of $Disc_i$ and the effect of deletion of items on point bi-serial correlations, test reliability and factorial validity.

## 5. REFERENCES

Avanoor V. and Mahendran P. (2018). Executive Function Rating Scale [EFRS]: A Study among Learning Disabled. Tool development: Item Generation and Item Analysis. *International Journal of Indian Psychology, Vol. 6,* (2), DIP: 18.01.031/20180602, DOI: 10.25215/0602.031

Chakrabartty, S.N. (2022). Disability and Quality of Life. *Health Science Journal*, Vol. 16, No. 12; 1 – 6

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

Chakrabartty, S.N. (2022). Measurements in mental tests through person space. *Current Psychology*, 41 (1). DOI: 10.1007/s12144-020-01033-3

Chakrabartty, S.N. (2020). Discriminating Value of Item and Test. *International Journal of Applied Mathematics and Statistics*, 59(3), 61 - 78

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 / bm- 1214182613

Denga, D.I. (1987). *Educational measurement, continuous assessment and psychological testing.* Calabar Rapid Educational Publishers.

Erhart M, Hagquist C, Auquier P, Rajmil L, Power M, Ravens-Sieberer U; European KIDSCREEN Group (2010). A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of a Quality of Life scale for children and adolescents. *Child Care Health Dev*. 36(4):473-84. doi: 10.1111/j.1365-2214.2009.00998.x.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement,*37(4), 827–838. https://doi.org/10.1177/001316447703700403

Feltz C.J, Miller G.E. (1996). An asymptotic test for the equality of coefficients of variation from k populations. *Stat Med*. 15(6):646-58. doi: 10.1002/(sici)1097-0258(19960330) 15:6<647: aid-sim184>3.0.co;2-p.

Ferrando, P. J. (2012): Assessing the discriminating power of item and test scores in the linear factor-analysis model. *Psicológica, 33*(1), 111-134.

Hankins M. (2007). Questionnaire discrimination: (re)-introducing coefficient Delta. *BMC Medical Research Methodology*. 7(1):19. doi: 10.1186/1471-2288-7-19.

Huang, Rui-Ting, Yu, Chung-Long,Tang, Tzy-Wen and Chang, Sheng-Chun (2021). A study of the use of mobile learning technology in Taiwan for language learning, Innovations in Education and Teaching International, 58:1, 59-71. DOI: 10.1080/14703297.2019.1628798

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

Krishnamoorthy, K., & Lee, M. (2013). Improved tests for the equality of normal coefficients of variation. *Computational Statistics*, 29 (1 – 2), 215 – 232. DOI:10.1007/s00180-013-0445-2

Lorenzo-Seva, U. and Ferrando,P. J.(2021): MSA: The Forgotten Index for Identifying Inappropriate Items Before Computing Exploratory Item Factor Analysis. *Methodology*, 17(4), 296–306 https://doi.org/10.5964/meth.7185

Marwick, B. and Krishnamoorthy, K. (2019): CV equality: Tests for the Equality of Coefficients of Variation from Multiple Groups. R software package version 0.1.3. https://github.com/benmarwick/cvequality

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*; 47, 149–174. doi: 10.1007/BF02296272

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

Muñiz, J., & Fonseca-Pedrero, E. (2019): Diez pasos para la construcción de un test. *Psicothema, 31*(1), 7-16. https://doi.org/10.7334/psicothema2018.291

Ogasawara, H. (2006): Approximations to the distribution of the sample coefficient alpha under non-normality. *Behaviormetrika*; 33(1), 3–26. https://doi.org/10.2333/bhmk.33.3

Parkerson HA, Noel M, Pagé MG, Fuss S, Katz J, Asmundson GJ (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*. Boston, MA: Allyn & Bacon.

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. 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 (4):201-204. DOI: 10.4103/2395-2296.189670

Raykov, T. (2008). Alpha if item deleted: a note on loss of criterion validity in scale development if maximizing coefficient alpha. *Br. J. Math. Stat. Psychol.*61, 275–285. doi: 10.1348/000711007X188520

Raykov, T. (2007). Reliability if deleted, not "alpha if deleted": Evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology*, 60, 201-216.

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. *Polygons*, 2, 1 – 35.

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,* 35(2), 67-71.

Sokal, R.R. and Rohlf, F.J. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research (3rd Ed.),* W.H. Freeman and Co., New York

Ten Berge J.M.F. & Hofstee W.K. (1999). Coefficient alpha and reliabilities of unrotated and rotated components. *Psychometrika*, 64(1):83–90. doi:10.1007/BF02294321