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

Evaluation of Psychometric Properties of Tests used in Educational Research

Satyendra Nath CHAKRABARTTY[1]

[1] *Indian Ports Association, HR Deptartment, Kolkata, India* *0000-0002-7687-5044*

ARTICLE INFO

Article History
Received 17.07.2025
Received in revised form 09.12.2025
Accepted 07.01.2026
Article Type: Research Article

ABSTRACT

It is imperative to assess psychometric parameters of tests and items by methodologically sound approaches and finding relationships among the psychometric qualities. The paper describes methods of computations of psychometric parameters of MCQ test and items under classical test theory, finds relationships among theoretically defined test reliability with factorial validity (FV) and separately with discriminating and difficulty values of test and items considering the entire data. Finding r_(tt(theoretical)) from a single administration offers significant benefits like computation of test error variance; testingH_0: r_(tt(theoretical))=1; finding battery reliability, estimation of true scores and its confidence interval. Relationships of r_(tt(theoretical)) derived with FV of standardized scores (〖FV〗 _(Z-scores)), test discriminating value (〖Disc〗 _T), Cronbach alpha in terms of item discriminating value 〖(Disc〗 _i) and 〖Disc〗 _T, Item reliability as function of 〖Diff〗 _i and 〖Disc〗 _T, alpha using PCA results (α_PCA) with 〖FV〗 _(Z-scores). The point of inflection of the negatively slopped reliability - discriminating value curve gives 〖Disc〗 _T for which test reliability is optimal; Common point of 〖Disc〗 _iand 〖Diff〗 _i curves could be used in item deletions to improve test reliability.

Keywords:
Theoretical reliability, ıtem reliability, battery reliability, factorial validity, discriminating value, eigenvalues

1. Introduction

Various scales and tests used in Educational Research for assessment in different educational streams, especially at the end of academic sessions, learning outcomes, and for selection purposes, which are MCQ type tests containing multiple-choice questions where a credit of "1" is given if the item is correct answered correctly and "0" for incorrect answer or omitted item. In addition, Likert scales and rating are also used in educational research to assess intensity and direction of respondents' feelings, to reveal important insights. Each type of tool uses sum of item scores to get test scores. Thus, the generated data consist of count of responses in each category of an item and score of an individual as sum of such responses give at best rank order information. Evaluation of internal structure of the test/scale is difficult by summative scores of items (Crutzen & Peters, 2017) and such scoring can affect reliability, validity of the test and classification using cut-off points (McNeish and Wolf 2020).

National Education Policy (NEP) 2020, Government of India is a game-changer which emphasizes learning assessment, testing of skills like analytical, critical thinking, conceptual clarity etc. It is imperative to assess psychometric parameters of such tests and constituent items by methodologically sound approaches and also to find relationships among the psychometric qualities related to tests.

Researchers use different methods to find test reliability (r_{tt}) and none is in conformity to the definition as $r_{tt} = \frac{S_T^2}{S_X^2}$ where S_T^2 and S_X^2 denote respectively variance of true score and observed score. Test reliability in terms

[1] Corresponding author's address: Flatt 4B, Cleopatra, DC 258, Street No. 350, Action Area 1, New Town, Kolkata 700156, India
e-mail: chakrabarttysatyendra3139@gmail.com

Citation: Chakrabartty, S. N. (2026). Evaluation of psychometric properties of tests used in educational research. *International Journal of Psychology and Educational Studies*, *13*(1), 28-39. https://dx.doi.org/10.52380/ijpes.2026.13.1.1547

of Test-retest reliability ($r_{Test-retest}$), Split-half reliability (r_{gh}), coefficients like Cronbach's alpha (α), McDonald's Omega (ω), etc. differ from theoretical definition of reliability and produce different results.
While, the popular Cronbach's alpha indicates consistency of scale over time periods, over different forms, and raters (Miller et al., 2013), test-retest reliability requiring two administrations reflects stability of a test, value of which differs with changes in time-interval between the two administrations. Its assumption of unchanged true scores during the time gap is not tenable due to effects of practice, learning during the time gap and potential minor differences in testing situations, etc. There is no consensus on time gap. $r_{Test-retet}$ could be high even if there is poor or no agreements. For example, if *retest-score* = $\alpha \pm \beta(test\ score)$, there will be no agreement but $r_{Test-retet}$ will be very high and close to unity. Berchtold (2016) used agreement unlike Jelenchick et al. (2012) who focused on correlation for $r_{Test-retet}$.

Split-half reliability (r_{gh})is based on correlation between the two sub-tests obtained by dichotomizing the test scores in two parallel subtests (g-th and h-th) defined as unchanged true score of i-th individual i.e. T_{ig}= T_{ih} implying $\overline{X_g} = \overline{X_h}$ and $X_g^2 = X_h^2$ which are commonly tested by t-test and F-test respectively. Such tests may be invalidated by presence of heteroscedastic errors, non-satisfaction of normally distributed data. r_{gh} as correlation between the two parallel sub-tests, depends on methods to dichotomize the original test scores and in many cases, parallality of the sub-tests is assumed without verifications through statistical tests. There is no universally agreed way of splitting a test in parallel valves. Thus, different values of Split-half reliability are possible for a test and for the same sample.Inter-item reliability considers the arithmetic mean (AM) of inter-item correlations (Cohen & Swerdlik, 2010). However, addition of correlations are not meaningful. Thus, AM of correlations of items and test scores is not meaningful (Garcia, 2012). Sample-driven mean of correlations between items and test scores was not favoured (Field, 2003). Similarly, average factor-loadings is not a meaningful standalone metric.

Computation of Cronbach's alpha needs pre-checking of its assumptions like uni-dimensionality and tau-equivalence i.e. equal factor loadings of items, which is rare for tests being used in educational research (Pronk et al. 2022). In real life, assumptions of alpha are not complied and different factor loadings are realistic (Teo and Fan, 2013). Moreover, outliers in the data affects alpha. There exists high number of reported cases on misuse of Cronbach's alpha (Cho & Kim, 2015; Sijtsma, 2009; Sijtsma & Pfadt, 2021). Trizano-Hermosilla and Alarado (2016) used Monte Carlo simulation to compare different measures of reliability and found that McDonald's Omega (ω) defined as $\omega = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum \theta_i}$ performed better than alpha where for the i-th item, λ_i denotes factor loading and θ_i represents the error variance. Omega does not require tau-equivalence assumption but assumes a unidimensional factor model. Computation of omega requires undertaking of Confirmatory Factor Analysis (CFA) first. Thus, factors affecting fitting of CFA model also affect value of omega. For small sample size, fitting of CFA model is difficult (Gagne & Hancock, 2006) and estimate of ω may be biased (Edwards et al., 2021). If data fit is problematic, estimation of omega is problematic and "should not be used" for reliability estimation (McDonald, 2011). Omega performed worse than α for: small samples, smaller number of items (test length), low factor loadings and weak inter-item correlations (Orcan 2023). Most measures of test reliability fail to satisfy ideal standards in decision making (Charter, 2003). Use of standard error of measurement (SEM) instead of test reliability was preferred (Zimmerman, 2007; Tighe et al. 2010). Reliability of battery of tests like Common Aptitude test (CAT), Differential Aptitude Tests (DAT), Primary Mental Abilities (PMA), tests to measure learning outcomes, achievement tests, etc. are multi-dimensional and violate assumption of alpha. In addition, cumulative measurement error of the component tests gets increased for the battery.

Test validity is commonly expressed as r_{XY} where test score and criterion score are denoted by X and Y respectively. Different choice of criterion score, different distributions of X and Y and mismatches of constructs being measured by X and Y may distort value of r_{XY}. Question arises whether it is validity for X or Y? If r_{XY} is high, what is the need for the test? If a test measures more than one factor, r_{XY} is the validity of which factors? Can we have test validity irrespective of criterion scores? One solution is offered by Factorial validity (FV) as ratio of the highest factor loading to the sum of all the factor loadings (Lozano et al. 2008).

Discriminating power of i-th item and the test in terms of their values ($Disc_i$and $Disc_T$) and their relationships including relationships with test reliability and validity are not usually reported as quality of a test.

Responsiveness indicates ability of a test to measure accurately the changes in score across time for a person or a group of persons and may be interpreted as longitudinal validity (Mokkink et al. 2021).

The paper describes methods of computations of quality psychometric parameters of MCQ test and items under classical test theory (CTT), finds relationships among such psychometric parameters including relationships of theoretically defined test reliability with FV and separately with difficulty/discriminating value of items and tests based on the entire data.

Table 1: *Symbol/notation table*

Variables	Denoted by capital letters like: *X*, *T*, *E* respectively for observed score, True score, Error score
Values taken by a variable	Denoted by small letters like: x_i, t_i, e_i, etc. Individual scores in the sub-tests are: x_{ig}, x_{ih} and corresponding true scores: t_{ig}, t_{ih}; and error scores: e_{ig}, e_{ih}
Statistic	Mean of the test by $\bar{X}, \bar{T}, \bar{E}$; and mean for the sub-tests: $\overline{X_g}, \overline{X_h}$ etc. Sample variance by S_X^2, S_T^2, S_E^2; For the sub-tests: $S_{X_g}^2, S_{X_h}^2$ Variance of a battery of tests by $S_{T(Battery)}^2$ Population variance by σ_X^2, σ_T^2, etc. Sample correlation by r_{XT}, r_{TE}, r_{TE}, $r_{E_1E_2}$, r_{gh} Population correlation by ρ_{XX_g}, ρ_{XX_h}
Reliability	r_{tt}; Reliability of the *i*-th test by $r_{tt(i)}$ Theoretical reliability by $r_{tt(Theoretical)}$= $\frac{S_T^2}{S_X^2} = 1 - \frac{S_E^2}{S_X^2}$ Reliability of a battery of tests by $r_{tt(battery)}$
Vector	By capital letter in bold like: $\boldsymbol{X_g}, \boldsymbol{X_h}, \boldsymbol{E_g}, \boldsymbol{E_h}$, Identity vector as $\boldsymbol{I} = (1, 1, \dots\dots 1)^T$, Vector of weights as $\boldsymbol{W}$ Length of vectors by $\|X_g\|$, $\|X_h\|$, $\|I\|$
Matrix	By capital letter like$\boldsymbol{S_{k\times k}}$ for matrix with *k*-rows and *k*-columns
Estimates	Estimates of true scores ($\hat{T}$)

2. Reliability as Per Definition

For a test, CTT assumes observed score *X* = *T* + *E* where $\bar{E} = 0$; $r_{TE} = 0$ and $r_{E_1E_2} = 0$ Chakrabartty et al.(2024) dichotomized a test taken by *N*-individuals in two parallel sub-tests where $t_{ig} = t_{ih}$ and S_{e_g}=S_{e_h} and proposed finding theoretical reliability of a test.

Here, score of *i*-th subject $x_i = x_{ig} + x_{ih} \Rightarrow t_{ig} + e_{ig} + t_{ih} + e_{ih}$

$\Rightarrow x_{ig} - x_{ih} = e_{ig} - e_{ih}$

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

where $\|X_g\| = \sqrt{\sum x_{gi}^2}$ and $\|X_h\|$= $\sqrt{\sum x_{hi}^2}$ denote respectively length of the vectors $\boldsymbol{X_g}$and $\boldsymbol{X_h}$ representing scores of the *g*-th subtest and *h*-th subtest respectively, θ_{gh} is the angle between the vectors $\boldsymbol{X_g}$ and $\boldsymbol{X_h}$ and $\theta_{gh}^{(E)}$is the angle between the unknown vectors E_g and E_h.

θ_{gh} can be found as $Cos\theta_{gh}$= $\frac{X_g^T X_h}{\|X_g\| \, \|X_h\|}$

But $r_{E_g,E_h} = 0$ for two parallel tests. Thus,

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

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

Theoretical test reliabilit is $r_{tt(Theoretical)} = 1 - \frac{S_E^2}{S_X^2}$ (2)

Considering *g*-th and *h*-th tests are parallel, equation (1) and (2) can be further simplified as

$S_E^2 = \frac{2\|X_g\|^2 (1 - Cos\theta_{gh})}{N}$ and (3)

$$r_{tt(Theoretical)} = 1 - \frac{2\|X_g\|^2 (1 - Cos\theta_{gh})}{NS_X^2} \qquad (4)$$

Avoiding assumptions of alpha, theoretical reliability as per (4) better handles the outliers amongst the test items, when compared to Cronbach alpha i.e. addition or removal of items with extreme scores, gives similar results in $r_{tt(Theoretical)}$ than alpha.

Empirical verification:

$r_{tt(Theoretical)}$ depends heavily on procedure of dichotomizing the test to two parallel sub-tests. Chakrabartty et al. (2024) presented an iterative process of splitting a MCQ test minimizing absolute difference of g-th vector$(\boldsymbol{X_g})$ and h-th vector $(\boldsymbol{X_h})$ using Bayesian learning of the item scores ensuring $\overline{X_g} = \overline{X_h}$ and $S_g^2 \approx S_h^2$. Such dichotomization performed better than odd – even splitting or partitioning a set of integers > 0 into two partitions, by maximising the inner product of the two partitioned vectors using polynomial-time approximation algorithms (Karmarkar & Karp, 1982) or usual splitting a dataset into training and test datasets in Machine Learning approach. The MCQ test with 50 items, administered among 912 examinees had mean = 16.24 and sample variance =19.63. Resultant parallel subtests as per the proposed iterative process gave$\overline{X_g} - \overline{X_h} = 0$; $\|S_g^2 - S_h^2\| = 0.17$ and $r_{gh} = 0.38$. Based on equation (2), value of $r_{tt(Theoretical)}$ was 0.93 > Cronbach alpha 0.91> Split-half reliability (r_{gh})

Benefits of theoretical reliability:

Theoretical reliability is isomorphic to its definition and facilitates the following:

(i) Test error variance (S_E^2) by equation (3), and $S_T^2 = S_X^2 - S_E^2$ against observations of Webb et al.(2006); Rudner & Schafes (2002) that theoretical r_{tt} is not possible since true scores are unknown at individual levels and r_{tt} is not perfectly precise (Zimmerman, 2007).

(ii) Testing null hypothesis H_0: $r_{tt(theoretical)} = 1$ is equivalent to H_0: $\sigma_X^2 = \sigma_T^2$ for which the test statistic is $F = \frac{S_X^2}{S_T^2}$ and reject H_0 if $F > F_{\alpha,(N-1,N-1)}$.

(iii) Chakrabartty (2020) proposed reliability of a test battery $(r_{tt(battery)})$ when battery score is defined as $Y_i = \sum_{i=1}^{K} W_i X_i$ where W_i is positive weight assigned to the i-th constituent test (∀ i= 1, 2, ….,K) and $\sum W_i = 1$, as:

$$r_{tt(battery)} = \frac{\sum_{i=1}^{K} r_{tt(i)} W_i^2 S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2W_i W_j Cov(X_i, X_j)}{\sum_{i=1}^{K} W_i^2 S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2W_i W_j Cov(X_i, X_j)} \qquad (5)$$

where $r_{tt(i)}$ and $S_{X_i}^2$denotes respectively reliability and variance of the i-th test.

Here, $var(Y) = \sum_{i=1}^{K} W_i^2 var(X_i)$ and S_T^2 of the battery can be evaluated by

$$S_{T(Battery)}^2 = \sum_{i=1}^{K} r_{tt(i)} S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2Cov(X_i, X_j) \qquad (6)$$

Instead of weighted sum, if battery score is computed without weights as $\sum_{j=1}^{K} X_j$ where X_j denotes score of the j-th test, battery reliability is obtained as

$$r_{tt(battery)} = \frac{\sum_{i=1}^{K} r_{tt(i)} S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2Cov(X_i, X_j)}{\sum_{i=1}^{K} S_{Xi}^2 + \sum_{i=1,i\neq j}^{K} \sum_{j=1}^{K} 2Cov(X_i, X_j)} \qquad (7)$$

But addition of scores of independent tests or tests with different degree of correlations is not meaningful and may be difficult to interpret. To avoid the problems of interpretation, Streiner et al. (2003) suggested to transform the total score to T-score as $T = \overline{X^*} + SD^*$ where $\overline{X^*}$ is the desired mean (say 50) and SD^* is the desired standard deviation (SD) (say10). Shahar (2017) found optimal weights to minimize variance of weighted average by Lagrange multipliers and Cauchy-Schwarz inequality. Chakrabartty, (2020) proved that positive weight vector $\boldsymbol{W}$satisfying $\sum_{i=1}^{k} W_i = 1$ minimizes variance $(Y) = W^T S W$ when $\boldsymbol{W} = \frac{1}{2}[\lambda S^{-1} e]$ where

$\boldsymbol{S_{k\times k}}$ denotes variance–covariance matrix of k-items of the test and $\lambda$is the Lagrangian multiplier computed as $\lambda = \frac{2}{e^T S^{-1} e}$

(iv) Estimation of True scores and Confidence Interval:

Chakrabartty (2022) proposed estimates of true scores of individuals ($\hat{T}$) for a given observed score X, as a linear regression i.e. $\hat{T}= \alpha + \beta X+ \epsilon$

$$\text{or } \hat{T} = \bar{X}(1-\beta) + \beta X + \epsilon = \bar{X} + r_{tt}(X - \bar{X}) + \epsilon \qquad (8)$$

since the regression coefficient $\beta = r_{XT}\frac{S_T}{S_X}$ and$\alpha = \bar{X}(1-\beta)$.

However, presence of heteroscedastic errors may make the standard errors calculated by OLS are biased and inconsistent, which invalidates standard hypothesis tests (like t-test, F-test) and estimation of confidence intervals, leading to incorrect inferences

Properties:

Properties satisfied by $\hat{T}$ given in equation (8) are:

- Mean of $\hat{T} = \bar{T} = \bar{X}$
- $Var\,(\hat{T}) < Var(T) \Rightarrow \hat{T}$ is more homogeneous than the T
- Variance of error on true score estimation (${S_\epsilon}^2$) < error variance of the test (${S_E}^2$)
- $r_{T\hat{T}} > r_{tt}$
- Confidence interval of a true score corresponding to observed score (x_0) is

$$\hat{T} \pm t_{N-2} S_\epsilon \sqrt{\frac{1}{N} + \frac{(x_0 - \bar{X})^2}{(N-1)S_X^2}} \qquad (9)$$

  where SD of residual is (S_ϵ)
- Confidence interval of test reliability is

$$r_{tt} \pm t_{\alpha/2,(N-2)}\left[\frac{\sqrt{\sum(t_i - \hat{T})^2}}{\sqrt{N-2}\sqrt{\sum(x_i - \bar{X})^2}}\right] \qquad (10)$$

(v) Tests of Parallelism:

Simultaneous testing of $\overline{X_g} = \overline{X_h}$ and $S^2_{X_g} = S^2_{X_h}$ for parallelism of two sub-tests is equivalent to testing goodness of fit of linear regression $(X_g - X_h) = \alpha + \beta((X_g + X_h)$. García-Pérez, (2013) showed that if X_g and X_h follow bivariate normal distribution, $F = \frac{(\sum {D_i}^2 - SSE)/2}{SSE/(n-2)} \sim F_{2,(N-2)}$ where SSE denotes sum of squares of residuals for linear regression of D on S. Other ways to test parallelism of two tests, suggested by Chakrabartty (2022) are as follows:

* Fit regression equations of $X = \alpha_1 + \beta_1 X_g$and also $X = \alpha_2 + \beta_2 X_h$ and then use likelihood ratio test of equal parameter models.

* Test significance of $\frac{\text{Mean sum of squares } (mss) \text{ due to deviation from the hypothesis}}{\text{Residual due to separate regression along}}$ with corresponding degrees of freedom.

* Testing equality of two correlations i.e. $H_0: \rho_{XX_g} = \rho_{XX_h}$by transforming the correlation coefficient to Z-scores using r-to-z transformation given by Fisher and statistical resting of significance by the observed Z-statistic.

* Each of g-th and h-th subtest consisting of $\frac{n}{2}$ items are parallel $\Leftrightarrow$ $Cos\ \beta_{Xg} = Cos\ \beta_{Xh} \Rightarrow \beta_{Xg} = \beta_{Xh}$ where β_{Xg} is the angle between $\boldsymbol{X_g} = (X_{1g}, X_{2g}, \dots \dots ., X_{\frac{n}{2}g})^T$and identity vector$\boldsymbol{I} = (1, 1, \dots \dots 1)^T$.

The angle between $\boldsymbol{X_h}$and $\boldsymbol{I}$ (β_{Xh}) is defined similarly. Here, $Cos\beta_{Xg}$=$\frac{\sum X_{ig}}{\|X_g\|\sqrt{N}}$ as $\|\boldsymbol{I}\| = \frac{n\sqrt{N}}{2}$ and $Cos\beta_{Xh}$= $\frac{\sum X_{ih}}{\|X_h\|\sqrt{N}}$

For parallel subtests (g-th and h-th),

$$\overline{X_g} = \overline{X_h} \Rightarrow \|X_g\|\ Cos\beta_{Xg} = \|X_h\|\ Cos\beta_{Xh} \text{ or } \frac{\|X_g\|}{\|X_h\|} = \frac{Cos\beta_{Xh}}{Cos\beta_{Xg}} \text{ and}$$

$$S^2_{Xg} = \frac{\|X_g\|^2}{N} - \overline{X_g}^2 = S^2_{Xh} = \frac{\|X_h\|^2}{N} - \overline{X_h}^2 \text{ which implies } \|X_g\|^2 = \|X_h\|^2$$

Thus, for parallel tests, $\|X_g\|^2 = \|X_h\|^2 \Rightarrow Cos\ \theta_{Xg} = Cos\ \theta_{Xh} \Rightarrow \theta_{Xg} = \theta_{Xh}$

Spruill, (2007) found distribution of dot product of two vectors (***X.Y***) where

$\|X\| = \|Y\| = 1$, can be approximated by Normal distribution for large sample. The method based on Cosine similarity without assuming distribution of_X_g and X_h may be used to assess whether two subtests are parallel.

Improving test reliability:

Reliability of MCQ test can be improved by deleting items which are rather ineffective in terms of difficulty values($Diff_i$) (commonly denoted by p_i) and discriminating values ($Disc_i$). However, computing $Diff_i$ considering entire data and $Disc_i$ ignoring 46% of the data are methodological inconsistent and interpretations of relationship between $Diff_i$ and $Disc_i$are difficult. For example, empirical observation of $r_{Diff_i,Disc_i}$ = 0.56 by Rao, et al. (2016) contradicts usual idea that very easy items (high $Diff_i$ values) and very difficult items (low $Diff_i$ values) give rise to poor discriminating value of items. Sim and Rasiah (2006) found $r_{Diff_i,Disc_i} > 0$ when $0.8 \leq Diff_i \leq 1.0$ and

< 0 when $0 \leq Diff_i \leq 0.20$ and a dome-shaped relationship considering all the items. Further investigations on $r_{Diff_i,Disc_i}$was felt needed (Chauhan, et al. 2013) along with effect of deleted items on test reliability and difficulty value ($Diff_T$) and $Disc_T$. Relationship between $Diff_i$ and $Disc_i$ and their associations with reliability, validity are also needed to know effect of deletion of items on test parameters including item-total correlations by point bi-serial correlations (r_{pbs}), $Disc_T$ or $Diff_T$.

Computation of item statistics:

Chakrabartty (2023) proposed following item statistics based on the entire data of a test with m-items taken by n-persons:

$$Diff_i = \frac{Number\ of\ correct\ answer\ to\ the\ i-th\ item}{n} = \frac{k}{n} \quad (11)$$

$$Diff_T = \frac{\bar{X}}{m} \quad (12)$$

Clearly, $0 \leq Diff_i \leq 1$ and $0 \leq Diff_T \leq 1$. Higher the $Diff_i$, easier the item is and higher the $Diff_{T,}$ easier is the test (high scoring).

$Diff_i \sim$ Binomial (n, p_i) with mean np_i, variance $np_i(1 - p_i)$.

$$Disc_i = \frac{S_{X_i}}{\bar{X_i}} = \sqrt{\frac{n-k}{nk}} = \text{Coefficient of variation of the item } (CV_i) \quad (13)$$

$$Disc_T = \frac{S_X}{\bar{X}} = \text{CV of the test } (CV_T) \quad (14)$$

Here, $0 \leq Disc_i < 1$ unlike [-1, +1] by usual method based on top 27% + bottom 27% of data.

$k = 1 \Rightarrow Disc_i$ is maximum and $k = (n-1) \Rightarrow Disc_i$ is minimum.

Relationship of item statistics with reliability:

Chakrabartty (2023) derived the following non-linear relationships:

$$Disc_i^{\ 2} = \frac{1-Diff_i}{n.Diff_i} = \frac{1-Diff_i}{k} \quad (15)$$

$$r_{tt(theoretical)} * Disc_T^{\ 2} = (\frac{S_T}{\bar{X}})^2 \quad (16)$$

$$Diff_T . Disc_T = \frac{S_X}{m} \tag{17}$$

$$r_{tt}\ (Disc_T)^2 = \frac{{S_T}^2}{{S_X}^2}\frac{{S_X}^2}{\bar{X}^2} = (\frac{S_T}{\bar{X}})^2 = (\frac{S_T}{\bar{T}})^2 \quad \text{since } \bar{X} = \bar{T} \tag{18}$$

$$\text{Cronbach } \alpha = \frac{m}{m-1}(1 - \frac{\sum_{i=1}^{m}\overline{X_\iota}^2 Disc_i^2}{\bar{X}^2 Disc_T^2}) \tag{19}$$

Equation (16) depicting non-linear negative relationship between $r_{tt(theoretical)}$ and $Disc_T$ indicates that increasing both r_{tt} and $Disc_T$ simultaneously is not possible. Equation (18) shows that product of $r_{tt(theoretical)}$ and $(Disc_T)^2$is equal to $CV^2_{True\ scores}$. Each of equations (15 – 19) holds even after deletion of one or more items, despite changes in S_X^2 and $Disc_T$ due to item deletions.

Criterion of Item deletion:

As per (15), $Disc_i$ changes with k. $Diff_i$ = $Disc_i$ at the point (k_0) where increasing curve of percentage$Diff_i$ and decreasing curve of percentage$Disc_i$ intersect. Empirically, Chakrabartty (2023) found that two curves of a MCQ test containing 50 items cut at k_0= 368 as shown below:

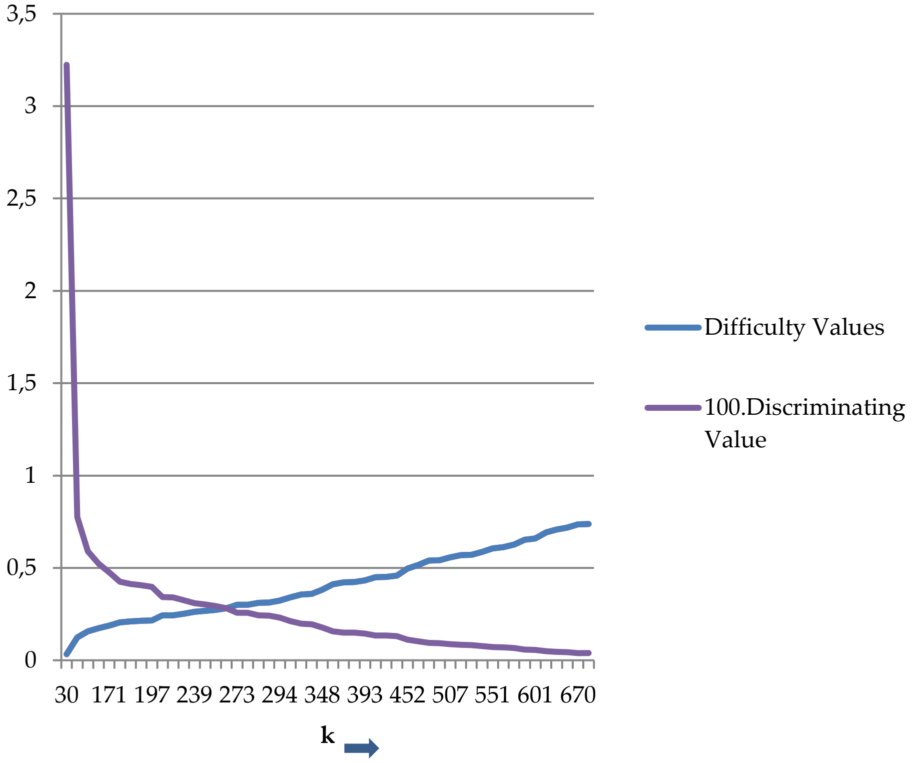

Figure 1: *Item-wise percentage $Diff_i$ and percentage $Disc_i$ curves*

At the point of intersection k_0= 368, difference between item difficulty (0.40395) and item discriminating (0.40245) was 0.00149. Right shift of k_0 implies increase in proportion of items with high $Diff_i$ (and low$Disc_i$). Value of k_0 is the integer solution of $\sqrt{\frac{n-k}{nk}} = \frac{k}{n}$

$$\text{or } k^3 = n(n-k) \tag{20}$$

Clearly, deletion of items will change values of $Diff_T$ & $Disc_T$.

Item reliability:

Reliability of an item of MCQ test (dichotomous variable) is commonly taken as point-biserial correlation (r_{pbs}) with test score (continuous variable) as

$$r_{pbs(i)} = \frac{(M_{pi}-M_{qi})\sqrt{Diff_i(1-Diff_i)}}{\bar{X}\ Disc_T} \tag{21}$$

where M_{pi}: Mean of test score of persons passing the i-th item and $M_{pi} + M_{qi} = 1$.

Clearly, $r_{pbs(i)}$and $Disc_T$ are negatively related. High value of $r_{pbs(i)}$ indicates that subjects passing the i-th item, also did well in the test. Clearly, $r_{pbs(i)} \geq 0 \Leftrightarrow (M_{pi} \geq M_{qi})$. Items for which $r_{pbs(i)}$ is very small or negative may be deleted or revised.

Correlation between two items:

Phi coefficient (Φ_{st}) reflects association between s-th and t-th item (each dichotomous) and is given by $\Phi_{st} = \frac{k_{11}k_{00}-k_{10}k_{01}}{\sqrt{k_{1.}k_{0.}k_{.0}k_{.1}}}$ where the symbols are clarified in the in the following 2 X 2 contingency table:

Items	*s* = 1	*s* = 0	Total
t = 1	k_{11}	k_{10}	$k_{1.}$
t= 0	k_{01}	k_{00}	$k_{0.}$
Total	$k_{.1}$	$k_{.0}$	*N*

Here, $\|X_t\|^2 = k_{1.}$ = number of subjects who passed the t^{th} item.

$\|X_s\|^2 = k_{.1}$ = number of subjects who passed the s^{th} item

$k_{11} = \sum\sum X_{is}X_{jt} = {X_s}^{\mathrm{T}}X_t$ denotes number of subjects passing both the s^{th} and t^{th} items.

$k_{00} = N - \sum\sum X_{is}X_{jt}$ denotes number of subjects failing both the s^{th} and t^{th} items.

Clearly,

$k_{10} = k_{1.} - k_{11}$= $\|X_t\|^2 - \sum\sum X_{is}X_{jt}$. Similarly, $k_{01} = k_{.1} - k_{11} = \|X_s\|^2 - \sum\sum X_{is}X_{jt}$ and $k_{00} = k_{0.} - k_{01} = (N - k_{1.}) - k_{01} = (N - \|X_t\|^2) - \|X_s\|^2 + \sum\sum X_{is}X_{jt}$

Thus, Φ_{st} can be expressed as function of $\|X_t\|^2$, $\|X_s\|^2$, ${X_s}^{\mathrm{T}}X_t$ and N as:

$$\Phi_{st} = \frac{{X_s}^T X_t(N-{X_s}^T X_t)-({X_t}^2-{X_s}^T X_t)(\|X_s\|^2-{X_s}^T X_t)}{\sqrt{\|X_t\|^2\|X_s\|^2(N-\|X_t\|^2)(N-\|X_s\|^2)}} \tag{22}$$

3. Validity

Validity of a test reflects the extent to which the test is able to measure the envisaged latent trait(s). To find test validity, researchers used different approaches like: CFA, Exploratory factor analyses (EFA), Structural equation modeling (SEM), etc. and found several factors could be measured by a test. Validity of MCQ was investigated based on difficulty and discrimination values of items, both measured in traditional ways of ignoring good portion of data (Patil et al. 2022) who found MCQs were reliable but not valid in medical education. Considering differential performance on responses and MCQ version of a test at a single time-point with 23 items, Ali et al. (2016) assessed validity by Cohen's $d = \frac{Mean_{\ Experimental\ group} - Mean_{Control\ group}}{SD_{Control\ group}}$

Construct validity:

Usual way to find construct validity is by correlation between test score (X) and criterion scores (Y) or express validity as the beta coefficient of regression Y on X. The latter require checking normality of residuals. For example, if X: 1, 2, 3, ….30 and $Y = X^2$, correlation between Y and $X^2 > 0.91$, But, regression of Y on X or X on Y are not justified despite high value of correlation since residuals do not support normal distribution and homoscedasticity (same variance), which are key assumption of Ordinary Least Squares regression, ensuring valid statistical inference. Problems of construct validity of a multidimensional test can be avoided by FV taken as: $FV = \frac{\lambda_1}{\sum \lambda_i}$ (23)

where λ_1 is the first (maximum) eigenvalue. FV explaining $\frac{\lambda_1}{\sum \lambda_i} \times 100\%$ of overall variability indicates validity of the main factor of the multidimensional test. Such FV avoids two administrations of the test and selection of criterion scale with similar factor structures (Parkerson et al. 2013). Significance of the largest eigenvalue can be tested by Tracy–Widom (TW) test statistic $U = \frac{\lambda_1}{\sum \lambda_i}$ which follows a TW-distribution i.e., distribution of the normalized $\lambda_1$of a Hermitian matrix for which each eigenvalue is real (Nadler, 2011). Extension of the concept of FV to assess factorial validity of battery of tests is suggested as a future study.

Reliability and Factorial validity:

Ten Berge and Hofstee (1999) suggested to find reliability of a test containing m-items as a function of λ_1 as

$$\alpha_{PCA} = (\frac{m}{m-1})\ (1 - \frac{1}{\lambda_1}) \qquad (24)$$

Clearly, λ_1 gives maximum value of α_{PCA} which is invariant under linear combination of scores of all the items of the test.

$$\text{Now, FV} = \frac{\lambda_1}{\sum \lambda_i} = \frac{\lambda_1}{Traces\ of\ the\ variance-covariance\ matrix} = \frac{\lambda_1}{\sum S^2_{X_i}}$$

For standardized item scores, $FV_{Z-scores}$ of a test containing m-number of items is $\frac{\lambda_1}{m}$ and

$$r_{tt(theoretical)} = \frac{S^2_T}{S^2_X} = \frac{S^2_T}{\sum \lambda_i + 2\sum^m_{i \neq j=1} Cov(X_i, X_j)} = \frac{S^2_T}{\frac{\lambda_1}{FV} + 2\sum^m_{i \neq j=1} Cov(X_i, X_j)} \qquad (25)$$

Equation (25) shows non-linear relationship between $r_{tt(theoretical)}$ and $FV_{Z-scores}$ where each term is computed from data and S^2_T may be estimated by (S^2_X - S^2_E) (from equation 3).

Relationship between FV and reliability as per α_{PCA} as given in equation (23) is:

$$\alpha_{PCA} = (\frac{m}{m-1})\ (1 - \frac{1}{\lambda_1}) = (\frac{m}{m-1})\ (1 - \frac{1}{FV.\sum \lambda_i}) = (\frac{m}{m-1})\ (1 - \frac{1}{m.FV_{Z-scores}}) \qquad (25)$$

The equation (25) indicates higher value of $FV_{Z-scores}$ increases α_{PCA}

4. Discussions

The paper describes evaluation of psychometric properties of tests and items used in educational research in a comprehensive manner. Major advantages include among others:

- $r_{tt(theoretical)}$ of a test requiring single administration offers significant benefits like computation of error variance of test; testing H_0: $r_{tt(theoretical)} = 1$; finding battery reliability and true score variance; estimation of true scores and confidence interval for an observed score (x_0), confidence interval of test reliability, etc.
- Inter-item reliability as average of item - test correlations has methodological limitations since addition of correlations is not meaningful. Instead, $r_{pbs(i)}$ could be taken as a better measure of item reliability.
- Use of the first eigenvalue (λ_1) for deriving FV, Cronbach's alpha (α_{PCA}) of a test and relationship between $r_{tt(theoretical)}$ and FV of standardized scores.
- Concept of discriminating value of test ($Disc_T$) and item ($Disc_i$) using the entire data. and their relationships with $Diff_T$ and $Diff_i$ including non-linear negative relationships between $r_{tt-theoretical})$ and $Disc_T$, expressing Cronbach alpha as function of $Disc_i\ and\ Disc_T$ and Item reliability by r_{pbs} as function of $Diff_i\ and\ Disc_T$.
- The point of inflection of the negatively slopped reliability- discriminating curve may be taken as the test discriminating value for which test reliability is optimal, since for a test, both reliability and discriminating value cannot be increased simultaneously. The point at which the $Disc_i$ and $Diff_i$ curves intersects could be used for deletion of items to improve test reliability.

- A number of statistical tests to test whether two tests are parallel.

The proposed approach of evaluation of psychometric properties of tests used to assess learning outcomes is well applicable also for various e-learning processes like flipped classroom (Osman et al. 2014), blended learning (Dos, 2014), synchronous and asynchronous learning (Chao & Chen, 2009) which focus on student-centered learning than teacher-centered learning to (Bergmann & Sams, 2012).

5. Conclusions

The approaches given in the paper with wide application areas help significantly in evaluation of better measures of reliability, validity, discriminating value of test and items using the entire data and derived relationships including relationship between reliability and validity, each as a function of largest eigenvalue. The relationships of reliability with discriminating value, validity etc. may help to maximize one parameter (say reliability) for a chosen value of another parameter. Future empirical investigations may explore such potentials, comparing power of various statistical tests proposed to test parallelism of two or more sub-tests and extension of factorial validity to battery of tests and construction of psychometric quality index of test and battery.

6. References

Ali, S. H., Carr, P. A., & Ruit, K. G. (2016). Validity and reliability of scores obtained on multiple-choice questions: Why functioning distractors matter. *Journal of the Scholarship of Teaching and Learning*, *16*(1), 1-14. https://doi.org/10.14434/josotl.v16i1.19106

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations* 9, 1–7. https://doi.org/10.1177/2059799116672875

Bergmann, J. & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education. https://escolaecofeliz.files.wordpress.com/2015/12/flip-your-classroom.pdf

Chakrabartty, S. N., Kangrui, W., & Chakrabarty, D. (2024). Reliable uncertainties of tests and surveys–a data-driven approach. *International Journal of Metrology and Quality Engineering*, *15*(4), 1-14. https://doi.org/10.1051/ijmqe/2023018

Chakrabartty, S. (2023). Item deletions based on difficulty values and discriminating values. *Edukasiana: Jurnal Inovasi Pendidikan*, *2*(4), 285-293. https://doi.org/10.56916/ejip.v2i4.455

Chakrabartty, S. N. (2022). Estimation of True Scores, True score variance, Reliability and tests of parallelism. *Journal of Modern Applied Statistical Methods*, *21*(2). https://doi.org/10.56801/Jmasm.V21.i2.3

Chakrabartty, S. N. (2020). Reliability of test battery. *Methodological Innovations*, *13*(2), 1-8. https://doi.org/10.1177/2059799120918340

Chakrabartty, S. N. (2018). Better composite environmental performance index. *Interdisciplinary Environmental Review*, *19*(2), 139-152. https://doi.org/10.1504/IER.2018.10014578

Chao, R. J. & Chen, Y. H. (2009). Evaluation of the criteria and effectiveness of distance e-learning with consistent fuzzy preference relations. *Expert Systems with Applications*, *36*(7), 10657-10662. https://doi.org/10.1016/j.eswa.2009.02.047

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *Journal of General Psychology*, *130*(3), 290–304. https://doi.org/10.1080/00221300309601160

Chauhan, P. R., Ratrhod, S. P., Chauhan, B. R., Chauhan, G. R., Adhvaryu, A., & Chauhan, A. P. (2013). Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot. *Biomirror*, *4*(6), 1-4.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well-known but poorly understood. *Organizational Research Methods, 18*(2), 207–230. https://doi.org/10.1177/1094428114555994

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

Crutzen, R. & Peters, G. J. Y. (2017). Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, 11, 242–247. https://doi.org/10.1080/17437199.2015.1124240

Dos, B. (2014). Developing and evaluating a blended learning course. *Anthropologist*, *17*(1), 121–128. https://doi.org/10.1080/09720073.2014.11891421

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics, 2,* 105–124. https://doi.org/10.1207/S15328031US020202

Gracia, 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

García-Pérez, M. A. (2013). Statistical criteria for parallel tests: A comparison of accuracy and power. *Behav Res.* 45, 999–1010. https://doi.org/10.3758/s13428-013-0328-z

Hojat, M., & Xu, G. (2004). A visitor's guide to effect sizes: Statistical significance versus practical (clinical) importance of research findings. *Advances in Health Sciences Education: Theory and Practice*, *9*(3), 241-249. https://doi.org/10.1023/B:AHSE.0000038173.00909.f6

Jelenchick, L. A., Becker, T., & Moreno, M. A. (2012). Assessing the psychometric properties of the Internet Addiction Test (IAT) in US college students. *Psychiatry research*, *196*(2-3), 296-301. https://doi.org/10.1016/j.psychres.2011.09.007

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4, 73–79. https://doi.org/10.1027/1614-2241.4.2.73

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287-2305. https://doi.org/10.3758/s13428-020-01398-0

Miller, M. D., Linn, R. L., & Gronlund, N. (2013). Reliability. In J. W. Johnston, L. Carlson, & P. D. Bennett (Eds.), *Measurement and assessment in teaching* (11th ed., pp. 108-137). Pearson.

Miller, M. D., Linn, R. L., & Gronlund, N. (2013). Validity. In J. W. Johnston, L. Carlson, & P. D. Bennett (Eds.), *Measurement and assessment in teaching* (11th ed., pp. 70-107). Pearson.

Mokkink, L., Terwee, C., & de Vet, H. (2021). Key concepts in clinical epidemiology: responsiveness, the longitudinal aspect of validity. *Journal of clinical epidemiology*, *140*, 159-162. https://doi.org/10.1016/j.jclinepi.2021.06.002

Nadler, B. (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. https://doi.org/10.1016/j.jmva.2010.10.005

Osman, S. Z., Jamaludin, R., & Mokhtar, N. E. (2014). Flipped classroom and traditional classroom: Lecturer and student perceptions between two learning cultures, a case study at malaysian polytechnic. *International Education Research* 2(4), 16–25. http://doi.org/10.12735/ier.v2i4p16

Parkerson, H. A., Noel, M., Pagé, M. G., Fuss, S., Katz, J., & Asmundson, G. J. (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

Patil, R. P., Bahekar, S. E., Kulkarni, M. D., & Baig, M. S. (2022). Evaluation of validity and reliability of multiple-choice questions in second MBBS competency-based medical education-based pharmacology examination of medical institute of India. *Int J Res Med Sci*, *10*, 2878.

Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, *29*(1), 44-54. https://doi.org/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. https://doi.org/10.4103/2395-2296.189670

Rudner, L. M. & Schafes, W. D. (2002): *Reliability*: *ERIC Digest*. www.ericdigest.org/2002-2/reliability/htm

Shahar, D. (2017). Minimizing the variance of a weighted average. *Open Journal of Statistics*, *7*, 216-224. https://doi.org/10.4236/ojs.2017.72017

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350.

Sim, Si–Mui & 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. https://doi.org/10.47102/annals-acadmedsg.V35N2p67

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K. & Pfadt, J. M. (2021). Part II: on the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, *86*, 843–860. https://doi.org/10.1007/s11336-021-09789-8

Spruill, M. C. (2007). Asymptotic distribution of coordinates on high dimensional spheres. *Electronic Communications in Probability*, *12*, 234–247. https://doi.org/10.1214/ECP.v12-1294

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use*. Oxford University Press..

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

Teo, T. & Fan, X. (2013). Coefficient alpha and beyond: issues and alternatives for educational research. *Asia Pac. Educ. Res. 22*, 209–213. https://doi.org/10.1007/s40299-013-0075-z

Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP (UK) examinations. *BMC Medical Education*, *10*(1), 40. https://doi.org/10.1186/1472-6920-10-40

Trizano-Hermosilla, I. & Alarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, 7, 769. https://doi.org/10.3389/fpsyg.2016.00769

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics*, 26, 81-124. https://doi.org/10.1016/S0169-7161(06)26004-8

Zimmerman, D. W. (2007). Correction for attenuation with biased reliability estimates and correlated errors in populations and samples. *Educational and Psychological Measurement*, *67*(6), 920–939. https://doi.org/10.1177/0013164406299132