Journal of Modern Applied Statistical Methods

Volume 21 | Issue 2 Article 3

Estimation of True Scores, True score variance, Reliability and tests of parallelism

Satyendra Nath Chakrabartty
Indian Statistical Institute, Indian Ports Association, Indian Maritime University,
chakrabarttysatyendra3139@gmail.com, ORCID: 0000-0002-7687-5044

Recommended Citation

Satyendra Nath Chakrabartty (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

Journal of Modern Applied Statistical Methods
2022, Vol. 21, No. 2,
Doi: 10.56801/Jmasm.V21.i2.3

ISSN 1538 – 9472

Estimation of True Scores, True score variance, Reliability and tests of parallelism

Satyendra Nath Chakrabartty
Indian Statistical Institute, Indian
Ports Association, Indian
Maritime University

Under classical test theory the paper presents a method to find test reliability (r_{tt}) as per theoretical definition from a single administration of the test, involving dichotomization of a test in parallel halves. The method helps to find point and interval estimations of test reliability, individual true score, true score variance, error variance and testing $H_o: r_{tt} = 1$. Dichotomization of a test in parallel halves requires simultaneous testing of equality of mean, variance and correlation for two parallel sub-tests (X_g and X_h) which can be done by testing (i) equality of regression line of observed score X on X_g and X on X_h (ii) equality of correlations between X and X_g and X and X_h, (iii) Normality of ($X_g - X_h$) or (iv) Cosine similarity (without assuming normal distribution of X_g and X_h). Reporting of theoretically defined reliability along with SD of true score/error score is recommended for a test.

Keywords: True score variance, Error variance, Reliability, Estimation and testing, Parallel tests.

1. Introduction

In classical test theory, observed test score of an individual is assumed to be sum of true score (error free score) and random error score. The additive model is Observed Score (X) = True Score (T) + Error Score (E) assuming error is randomly distributed around 0 i.e. average of error score ($\bar{E}$)= 0; correlation between true score and error score (r_{TE}) = 0; correlation between two series of error scores ($r_{E_1E_2}$) = 0. True score refers to the fraction of the score which is replicable or reliable. Statistic to summarize a phenomenon is concerned with how much of the statistic represents the true score and how much is error. Thus, it is critical to estimate accurately the true score component for each examinee along with variance of error scores (${S_E}^2$) and reliability (r_{tt}) as ratio of true score variance (${S_T}^2$) and observed score variance (${S_X}^2$). Standard deviation (SD) of errors of measurement (S_E) that are associated with test scores from a particular group of examinees is known as standard error of measurement (*SEM*), which reflects the extent of variation or spread in the measurement errors for a test and is frequently used to find bands around observed

test scores in the form $X \pm SEM$. Two important relationships involving error variance, true score variance and reliability are ${S_X}^2 = {S_T}^2 + {S_E}^2$ and $r_{tt} = \frac{{S_T}^2}{{S_X}^2} = 1 - \frac{{S_E}^2}{{S_X}^2}$. Thus, ${S_T}^2, {S_E}^2$ and r_{tt} are inter-related and proper estimation of any of these may help to find estimates of the other two since, by definition ${S_T}^2 = r_{tt}.{S_X}^2$.

Webb et al.(2006); Rudner & Schafes (2002) were of the view that reliability as per theoretical definition is not practical or impossible to compute since true scores of individuals taking the test are not known.. In addition, reliability coefficients are not perfectly precise (Zimmerman, 2007).The imprecision may be carried over into the estimation of ${S_T}^2, {S_E}^2$ and thus affect estimation of true scores of the examinees.

Meaningful comparison of groups demands error variance is invariant across groups. The assumption of constant error variance across individuals, irrespective of learning or other variables influencing the latent traits of the test may not be taken as a rule (Kline, 2005). Generalizability theory (G-theory) also makes similar assumption of constant error variance for all individuals with different true scores (Shavelson & Webb, 2012). Such assumption is not consistent with findings of Hedge et al. (2018) who demonstrated that test reliability decreases as within group variance increases. Williams et al. (2022) also found substantial individual variation in the error structure of cognitive tasks and hence different reliabilities for different groups. *SEM* is different at various score levels, and *SE* for the entire test does not adequately summarize the error propensity of most examinees (Feldt et al. 1985). Thus, *SEM* could be a test characteristic and also a score characteristic which varies within a group. Empirically, Lord (1959) determined that S_E is directly proportional to the square root of the number of items ($\sqrt{n}$) and found high $r_{S_E,\sqrt{n}}$ at the level of 0.99. However, the general relationship between S_E and $\sqrt{n}$ depends on type of reliability considered.

The existing methods of finding test reliability use a variety of ways and none of them is isomorphic to the definition of $r_{tt} = \frac{{S_T}^2}{{S_X}^2}$. Thus, there is potency for confusion over the trustworthiness of a test, emanating out of inconsistencies amongst different available methods to compute reliability coefficient.

Based on the general (congeneric) model for reliability, Cho, (2016) defined reliability of a test with *n*-items as $\frac{\sum_{i=1}^{n} {\lambda_i}^2}{{S_X}^2}$ where λ_i is the loading on the *i*-th item such that variance of the *i*-th item's score is $\lambda^2 + {S_{Ei}}^2$ where ${S_{Ei}}^2$denotes error variance of the *i*-th item. The model $r_{tt} = \frac{{S_T}^2}{{S_X}^2} = \frac{{S_X}^2 - {S_T}^2}{{S_X}^2}$ gets reduced to general congeneric if $\sum_{i=1}^{n} {\lambda_i}^2$ is replaced by ${S_X}^2 - {S_T}^2$.

Trafimow, (2014) found that estimate of true core variance ($\widehat{\sigma_T^2}$) of a test may be different for Experimental group and Control group even if r_{tt} is fixed and interpretation of low value of resulting error variance in one group (say experimental group) over another group (say control group) becomes problematic. In addition, test

reliability (r_{tt}) can be measured in different ways giving different values. Thus, true score variance as the product of observed score variance and test reliability would take different values depending on the method of computing test reliability.

Kristof, (1969) derived maximum-likelihood estimators of true score variance and error variance for mental tests under six different hypotheses of equivalent measurements. Estimation of true score variance using analysis of variance was dealt with by Jackson, (1973) where persons corresponds to "treatments" and scores on number of parallel tests to replications and observed need to have point estimation and interval estimation of true score from a single administration of a test. An empirical Bayes method for true score estimation involving a set of assumptions was suggested by Cressie, (1979).

Usual methods of computing test reliability aiming to indicate consistency, repeatability, precision, trustworthiness are test-retest reliability, parallel test reliability, split half method and the method of internal consistency. The test-retest method and the parallel method need two administrations of the test or parallel tests to the same sample and test reliability is computed by correlation between the two test scores, without making any assumption about the relationship between the items in the test or tau-equivalence or uncorrelated measurement errors of the items, etc. Test–retest reliability of a single test may have different values depending on time-gap between administrations. For parallel forms of the test, it is necessary to test that the two parallel forms are actually parallel ensuring equality of true score of *i*-th person in *g*-th sub-test and in *h*-th sub-test.

Reliability from single administration of the test is usually obtained as split-half reliability or internal consistency in terms of Cronbach alpha. The former considers dichotomization of the test score in parallel halves and reliability is expressed as correlation between two parallel sub-tests. However, split-half reliability is not unique and depends heavily on procedure of dichotomization ensuring that subtests are parallel. Frequent use of Cronbach alpha without checking the underlying assumptions have resulted in confusions regarding its proper use and interpretations (Schmitt, 1996; Cortina, 1993), Violation of assumptions like unidimensionality of the test, essentially tau-equivalent, etc. often lowers the value of alpha (Green, 2005; Graham et al. 2006). In practice, data satisfying all assumptions of alpha may not be viable (Teo and Fan, 2013). Cortina, (1993) found that Cronbach's alpha is a lower-bound estimate of reliability. Higher value of alpha may indicate higher measure of unidimensionality of the test but, alpha for multi-dimensional tests may be more than the same of one-dimensional tests (Cortina, 1993). Lord and Novick, (1968) showed that coefficient α equals the reliability if and only if the items in the tests are mutually essentially τ –equivalent and in all other cases, coefficient α is underestimated. Cronbach alpha is equivalent to Guttman's λ_3 which is ≤ Gutman's λ_2 and is not the best estimate (Guttman, 1945). Attempts to have better lower bounds culminated in the theory of the greatest lower bound (glb), discussed thoroughly by Ten Berge et al. (1981). However, computation of the glb is not simple and may be seriously biased. Alpha cannot simply be interpreted as an index for the internal consistency of a test (Green, 2005). Limitations of Cronbach's alpha

have been extensively addressed by researchers like Verhelst, (2000); Sijtsma, (2009); Ritter, (2010); Eisinga, et al. (2013); Panayides, (2013), etc. Thus, no popular method of finding reliability uses the theoretical definition of reliability resulting in different values of error variance and reliability for the same test even if the sample remains unchanged.

The paper proposes a method of computing test reliability as per theoretical definition from single administration along with point estimation and interval estimation of true score, true score variance, *SEM* and discuss properties of such estimates to see how these facilitate to have population estimates and testing of hypothesis.

The paper is organized as follows. Point estimation of true score for a given value of observed score using linear regression analysis and properties of such estimation are discussed in the following Section. This is followed by estimation of true score using theoretical definition of test reliability. The method of finding population estimates ${\sigma_X}^2, {\sigma_T}^2, {\sigma_E}^2$ and testing statistical hypothesis of population reliability is equal to one are expounded upon. Confidence interval of true score and test reliability are discussed in the following section along with prediction interval of future value of true score. Methods of testing parallelism are presented in next section. The paper is rounded up by recalling the salient outcomes of the work.

2. Point Estimation of True Score:

2.1 Using linear regression:

Consider an aptitude or achievement test consisting of n – items ("1" for correct answer and "0" otherwise) have been administered to *N* subjects. For a given observed score *X*, point estimation of the true score ($\hat{T}$) may be obtained as a linear regression on the observed score. The model is $\hat{T} = \alpha + \beta X + \epsilon$ where the regression coefficients are

$\beta = r_{XT}\frac{S_T}{S_X}$ and $\alpha = \bar{X}(1-\beta)$

Thus, the regression line is

$$\hat{T} = \bar{X}(1-\beta) + \beta X + \epsilon = \bar{X} + r_{tt}(X - \bar{X}) + \epsilon \quad (1)$$

where ϵ denotes the error in prediction of true score; clearly, $\bar{\epsilon} = 0$

The model represented in (1) helps to estimate true score of a subject as a linear function of his/her observed score using reliability of the test.

2.2 Observations:

As per the model,

i) Mean of estimated true scores i.e. mean of $\hat{T} = \bar{X} = \bar{T}$

ii) $Var(\hat{T}) = \beta^2(X) = {r_{tt}}^2(X)$ (2)

since $\frac{\text{Var }(\hat{T})}{Var(T)} = \frac{r_{tt}{}^2 \text{ Var}(X)}{r_{tt}Var(X)} = r_{tt} < 1$ which implies $Var\ (\hat{T}) < Var(T)$

Thus, distribution of $\hat{T}$ is more homogeneous than the same for T. If ${S_T}^2$ is replaced by${S_{\hat{T}}}^2$, test reliability will be less.

iii) *Proposition 1: Variance of error on estimation of true score* $({S_\epsilon}^2)$ *is less than the error variance of the test* $({S_E}^2)$

Proof: Here, the residual variance ${S_\epsilon}^2 = \frac{1}{N}\sum(\alpha + \beta X_i - T_i)^2 = {S_T}^2(1 - r_{tt})$

But, ${S_E}^2 = {S_X}^2(1 - r_{tt})$ by definition which implies $\frac{{S_E}^2}{{S_\epsilon}^2} = \frac{{S_X}^2}{{S_T}^2} = \frac{1}{r_{tt}} > 1$

In other words, variance of error in prediction of true score from a linear regression model is less than the test error variance ${S_E}^2$. This may be taken as goodness of the chosen model of estimating true score from a linear regression equation.

Clearly, higher value of reliability will result in lower value of ${S_\epsilon}^2$ and better estimates of $\hat{T}$

iv) *Proposition 2: Correlation between T and* $\hat{T}$ *as per the model is higher than test reliability*

Proof: Let Z_T and $Z_{\hat{T}}$ are the standardized variables obtained from T and $\hat{T}$.

Here, $Z_T = \frac{T-\bar{T}}{S_T}$ and $Z_{\hat{T}} = \frac{T-\bar{\hat{T}}}{S_{\hat{T}}}$ since mean of $\hat{T} = \bar{X} = \bar{T}$

$\Longrightarrow Z_{\hat{T}} = \frac{X-\bar{X}}{S_X}$ using (1) and (2) (3)

Now $Var(Z_T - Z_{\hat{T}}) = V(Z_T) + V(Z_{\hat{T}}) - 2Cov(Z_T, Z_{\hat{T}})$

$= 2[1 - Cov(Z_T, Z_{\hat{T}})]$ (4)

Here, $Cov(Z_T, Z_{\hat{T}}) = \frac{1}{N}\sum Z_{T_i}Z_{\hat{T}_i}$ since mean of Z_T = Mean of $Z_{\hat{T}} = 0$

$= \frac{1}{N}\sum(\frac{X_i-\bar{X}}{S_X})(\frac{T_i-\bar{T}}{S_T})$ using (3)

$= \frac{Cov(T,X)}{S_T S_X} = r_{XT} = \sqrt{r_{tt}} > r_{tt}$

It is well known that correlation between two standardized variables Z_X and Z_Y is given by $r = 1 - \frac{S^2_{(Z_X - Z_Y)}}{2}$ (Rodgers & Nicewander, 1988)

Accordingly, $r_{T\hat{T}} = r_{XT} = \sqrt{r_{tt}} > r_{tt}$

High correlation between T and $\hat{T}$ indicates another measure of goodness of the model chosen.

Point estimation of true scores will help to find frequency distribution of observed scores and same for estimated true scores and also for estimated error scores satisfying the following properties:

- $\bar{X} = \bar{T} = \bar{\hat{T}}$
- $Var\ (\hat{T}\ ) < Var(T)$
- Correlation between T and $\hat{T}$ exceeds reliability i. e. $r_{T\hat{T}} > r_{tt}$
- Mean of $\hat{E} = 0$
- $Var\ (E) > Var\ (\hat{E})$

Using $\hat{T}$ of each individual taking the test, one may undertake computation of the probability that the percentile true score of the *i*-th examinee is *t*, given the observed percentile score of the examinee is *x* and reliability is r_{tt} , i.e. Prob.($T \leq t|\ r_{tt} = r.\ X \leq x$).

The pertinent question is which reliability to be used in (1) to estimate true scores?

2.3 Reliability as per definition

Since higher value of reliability implies lower value of S_E^2 i.e. better prediction of true score, question arises regarding choice of reliability for estimating true score of a subject with knowledge of his/her observed score. However, no existing method of finding test reliability uses directly its definition as $r_{tt} = \frac{{S_T}^2}{{S_X}^2}$. Chakrabartty (2021) suggested a method of obtaining test reliability as per the theoretical definition along with computation of error variance and true score variance from single administration of the test. Suppose a test with *n*-items administered among *N*-persons is dichotomized into parallel halves in the form of *g*-th and the *h*-th subtests. Item-wise scores of the subtests can be represented respectively by vector $\boldsymbol{X_g} = (X_{g1}, X_{g2}, \ldots\ldots, X_{g^{n}/_2})^T$ and $\boldsymbol{X_h} = (X_{h1}, X_{h2}, \ldots\ldots, X_{h^{n}/_2})^T$ where lengths of the parallel sub-tests are $\|X_g\| = \sqrt{\sum {X_{gi}}^2}$ and $\|X_h\| = \sqrt{\sum {X_{hi}}^2}$. The cosine of the angle between the vectors $\boldsymbol{X_g}$ and $\boldsymbol{X_h}$ is given by $Cos\theta_{gh} = \frac{X_g^T X_h}{\|X_g\|\ \|X_h\|}$

Here, the error variance of the entire test is

$${S_E}^2 = \frac{1}{N}[\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\ \|X_h\| Cos\theta_{gh}] \quad (5)$$

and the test reliability as per theoretical definition is

$$r_{tt} = 1 - \frac{{S_E}^2}{{S_X}^2} = 1 - \frac{\|X_g\|^2 + \|X_h\|^2 - 2\|X_g\|\ \|X_h\| Cos\theta_{gh}}{N{S_X}^2} \quad (6)$$

In case of $\|X_g\| = \|X_h\|$ since $\boldsymbol{X_g}$ and $\boldsymbol{X_h}$ are parallel, (6) and (7) can be written respectively as

$$S_E^2 = \frac{2\|X_g\|^2}{N}(1 - Cos\ \theta_{gh}) \quad (7)$$

and $r_{tt} = 1 - \frac{2\|X_g\|^2}{NS_X^2}(1 - Cos\ \theta_{gh})$ (8)

2.4 Properties:

1. Equation (5) helps to find value of error variance of the test and hence true score variance ${S_T}^2$ as $(S_X^2 - S_E^2)$ and use them directly to find reliability of the test as per the theoretical definition of reliability from a single administration, in terms of length of score vectors of two parallel tests and angle between such vectors. Thus, it is possible to find true score variance from the data and to calculate reliability co-efficient that conforms to the theoretical definition even if true scores of individuals taking the test are not known.
2. Assuming normal distribution of *X*, computed value of sample S_E^2 may be used to find rough estimate of the true score of an examinee for a given observed score as $X \pm SEM$. Better will be to estimate of reliability from (6) and use it in (1) to find $\hat{T}$.
3. Relationship can be verified between $S_{\hat{T}}^2$ obtained as difference between S_X^2 and S_E^2 where S_E^2 is computed using (5) and $Var\ (\hat{T})$ from model (1) where r_{tt} is as per (6)
4. Value of correlation between two parallel sub-tests r_{gh}, as an estimate of Split-half reliability is different from theoretical value of r_{tt} obtained by (6). It can be proved that if the parallel tests satisfy equality of means and variances, the split-half reliability r_{gh} will be maximum.

Proposition 3: If splitting a test results in sub-tests g and h with $\overline{X_g} = \overline{X_h}$ *and* $S_g^2 = S_h^2$*, then split-half correlation* r_{gh} *is maximum.*

Proof: Let the regression line of X_g on X_h be $X_g = \alpha_1 + \beta_1 X_h$ where $\beta_1 = r_{gh}\frac{S_g}{S_h}$ and the regression line of X_h on X_g be $X_h = \alpha_2 + \beta_2 X_g$ where $\beta_2 = r_{gh}\frac{S_h}{S_g}$

Now, $S_g^2 = S_h^2 \implies \beta_1 = \beta_2$ this implies $\alpha_1 = \overline{X_g} - r_{gh}\overline{X_h}$ and $\alpha_2 = \overline{X_h} - r_{gh}\overline{X_g}$

Since $\overline{X_g} = \overline{X_h}$, it follows $\alpha_1 = \alpha_2 = \overline{X_g}(1 - r_{gh})$

Thus, the regression line of X_g on X_h and that of X_h on X_g are coincident with equal regression coefficients if $r_{gh} = 1$

Equivalently, departure from $r_{gh} = 1 \Leftrightarrow$ departure from $S_g^2 = S_h^2 \Leftrightarrow$departure from parallelism

5. It can be proved that using r_{gh} to estimate true scores will be in excess than r_{tt} obtained by using the classically defined reliability, for high values of the observed scores and under-estimated for low values of *X*.

Proposition 4: Let $\hat{T}_M$ *denotes prediction of true score for finding reliability. Then* $\hat{T}_{split-half} \geq \hat{T}_{Classical}$ *for* $X \geq \bar{X}$ *and* $\hat{T}_{split-half} \leq \hat{T}_{Classical}$ *for* $X \leq \bar{X}$

Proof: For $r_{gh} \geq r_{tt}, we\ have$

$$\hat{T}_{split-half} - \hat{T}_{Classical} = \left[\bar{X} + r_{gh}(X - \bar{X})\right] - [\bar{X} + r_{tt}(X - \bar{X})]$$
$$= (X - \bar{X})\,(r_{gh} - r_{tt}) \geq 0\ for\ X \geq \bar{X} \text{ and}$$
$$\leq 0\ for\ X \leq \bar{X}$$

It is suggested to use reliability as per (6) primarily because of its theoretical advantages without involving any assumptions of distributions of the observed or underlying variables.

2.5 Estimation and test of reliability

For a given data set, it is possible to find sample value of S_E^2, ${S_T}^2$, S_X^2, and r_{tt} using (5) and (6) and ${S_T}^2 = S_X^2 - S_E^2$. However, these values are likely to differ for different samples. Hence, it may be useful to have population estimates of these parameters. Unbiased and consistent estimate of variance of observed score is $\sigma_X^2 = \frac{1}{N-1}\sum(X_i - \bar{X})^2$ and can be written as $\frac{N}{N-1}S_X^2$. The estimate follows Gamma distribution with parameters $(N-1)$ and $\frac{(N-1)}{N}\sigma^2$ (Weisstein, 2003). Following similar approach, one can find σ_T^2 and σ_E^2 with the sample estimate of S_E^2 from (5) and ${S_T}^2$ as $(S_X^2 - S_E^2)$.

However, $\frac{\sigma_T^2}{\sigma_X^2}$ may not be a good estimate of population reliability since distribution of two correlated Gama variables will be too complex and beyond the scope of the paper. Confidence interval of test reliability is discussed at a later section.

Reliability as per equation (6) also helps to test whether the population reliability is equal to one. Since $r_{tt} = \frac{S_T^2}{S_X^2}$ as per the definition, the test is equivalent to testing $H_0: \sigma_X^2 = \sigma_T^2$ against $H_1: \sigma_X^2 > \sigma_T^2$ which can be tested using usual *F*-test where test statistic $F = \frac{S_X^2}{S_T^2}$ and reject H_0 if the test statistic *F* is too large i.e. if $F > F_{\alpha,(N-1,N-1)}$.

Error variance of the test computed from (5) or σ_E^2 may be mentioned along with Test reliability using (6) while reporting a test.

3. Confidence Interval:

3.1 Confidence Interval of True Score:

Confidence interval of true score represents a range of score that is likely to contain the true score for a given value of the observed score with a specified probability. Popular approach of confidence interval for true score corresponding to an observed score (X_0) using *SEM* of the test is given by $X_0 \pm Z_{95\%}S_E$ where $Z_{95\%}$ is the Z-score from a normal distribution table corresponding to a Z-score below which 95 % of the area of the standard normal distribution falls and S_E denotes *SD* of error of

measurement of the test. Thus, a confidence interval using S_E helps to estimate (with a certain level of confidence) underlying true score corresponding to a given value of the observed score.

Note that as test reliability increases (which means S_E^2 decreases), length of confidence intervals get narrower. Smaller confidence interval for any given Z % implies more accuracy. For $r_{tt} = 1,$ the observed score would equal the true score.

However, Leininger, (2013) suggested to find confidence interval of true score using *SD* of error of prediction i.e. *SD* of residual(S_ϵ). A confidence interval for $E(T/X_0)$, the expected value of Y for a given observed scoreX_0, is

$$\hat{T} \pm t_{N-2} S_\epsilon \sqrt{\frac{1}{N} + \frac{(X_0 - \bar{X})^2}{(N-1)S_X^2}} \tag{9}$$

Note that the width of the confidence interval for $E(T)$ increases as X_0 moves away from the center. In other words, as $(X_0 - \bar{X})^2$ increases the margin of error of the confidence interval. Conceptually, we are more certain of our predictions around the center of the data i.e. $\bar{X}$ than at the edges.

But to predict an interval of future values of T for a given X_0, Leininger, (2013) suggested prediction interval as

$$\hat{T} \pm t_{N-2} S_\epsilon \sqrt{1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{(N-1)S_X^2}} \tag{10}$$

The formula is similar to (9), except the variability is higher since there is an added 1 in the formula. If we repeat the study of obtaining a regression data set many times, each time forming a X% prediction interval at X_0 and see what the future value of T is at X_0 , then roughly X% of the prediction intervals will contain the corresponding actual value of T

The following may be noted:

i) A prediction interval is similar to a confidence interval, except that the prediction interval is designed to cover a "moving target", the random future value of T, while the confidence interval is designed to cover the "fixed target", the average (expected) value of T, $E(T)$, for a given X_0.

ii) Since, prediction intervals deal with the individual observations in a population as well as the parameter estimates, prediction intervals are wider than the confidence interval calculated for the same data set. Hence, prediction intervals are also more susceptible to the assumption of normality than are confidence intervals.

iii) The prediction interval takes into account tendency of $\hat{T}$ to fluctuate from its mean value, while the confidence interval simply needs to account for the uncertainty in estimating the mean value.

iv) For a given data set, the error in estimating true score increases as X_0 moves away from $\bar{X}$. In other words, length of both confidence and prediction intervals will increase with increase of $|X_0 - \bar{X}|$ or $(X_0 - \bar{X})^2$.

3.2 Confidence Interval of Test reliability:

For a given data set, we have estimated values of $r_{tt}, \bar{X}$ and S_ϵ which are used to find point estimates or interval estimates of true score. A different data set is likely to result in different values of the above. So, a need could be felt to have a measure of the accuracy of estimate, such as a confidence interval of test reliability. The model at (1) may be used to find such confidence interval of r_{tt}.

Under the assumptions of the simple linear regression model, a two sided, (1−α) 100% confidence interval for the slope parameter β suggested by Leininger, (2013) is:

$$b \pm t_{\frac{\alpha}{2},N-2} \left(\frac{\sqrt{N}\,\hat{\sigma}}{\sqrt{N-2}\sqrt{\sum(X_i-\bar{X})^2}}\right) \tag{11}$$

or equivalently: $\hat{\beta} \pm t_{\frac{\alpha}{2},N-2}\sqrt{\frac{MSE}{\sum(X_i-\bar{X})^2}}$

where $b = \hat{\beta} = \frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}$ and $\hat{\sigma}^2 = \frac{1}{N}\sum(Y_i - \hat{Y})^2$

Using the above in our model $\hat{T} = \alpha + \beta X$

where $\beta = r_{XT}\frac{S_T}{S_X} = r_{tt}$ and $\alpha = \bar{X}(1-\beta)$, confidence interval of the regression coefficient $\beta = r_{tt}$ is

$$r_{tt} \pm t_{\frac{\alpha}{2},N-2} \left(\frac{\sqrt{\sum(T_i-\hat{T})^2}}{\sqrt{N-2}\sqrt{\sum(X_i-\bar{X})^2}}\right) = r_{tt} \pm t_{\frac{\alpha}{2},N-2}\left(\frac{S_\epsilon}{\sqrt{N-2}S_X}\right) \tag{12}$$

The above could be interpreted as follows:

If large number of samples of N persons are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population slope i.e. r_{tt} is 1 – α.

3.3 Test of parallelism

The method of computing reliability as per definition involves dichotomization of a test in parallel halves. Thus, it is necessary to test that *g*-th and *h*-th subtests are parallel. The hypotheses of equality of mean, variance and correlation of parallel tests can be tested separately or simultaneously as a single multidimensional hypothesis. Testing parallelism of only two tests, involving a single correlation—can be treated with a simultaneous testing of equality of means and variances which is equivalent to testing for null slope and intercept in the regression of $D = X_g - X_h$ on $S = X_g + X_h$. Miguel & Garcia, (2013) recommended use of Bradley–Blackwood test having adequate power to detect differences in means or variances because of its simplicity and its minimally better performance. Under the assumption of bivariate normal distribution of X_g, X_h, the test statistic F =

$\frac{(\sum D_i^2 - SSE)/2}{SSE/(n-2)}$ is distributed *F* with 2 and $(n-2)$ degrees of freedom; where *SSE* is the residual sum of squares from the regression of *D* on *S*.

In addition, following methods are proposed for statistically assessing parallelism of two tests:

* Test equality of regression lines of *X* on X_g and *X* on X_h by ANOVA (Rao, 1952)

* Significance of the ratio of mean sum of squares due to deviation from the hypothesis to residual due to separate regression along with corresponding degrees of freedom may help to accept or reject the hypothesis.

* Testing equality of two correlations $X \& X_g$ and $X \& X_h$ i.e. $H_0: \rho_{XX_g} = \rho_{XX_h}$ using Fisher r-to-z transformation or by studentized permutation test for testing equality of correlation coefficients in two populations

* Test $Z = X_g - X_h$ follows Normal distribution by usual test of normality.

* Cosine similarity: If a test with *n*-items (assume *n* is even) is administered to *N*-individuals is dichotomizes to *g*-th and *h*-th parallel subtests, we get score vectors $\boldsymbol{X_g} = (X_{g1}, X_{g2}, \ldots\ldots, X_g{}_{n/2})^T$ and $\boldsymbol{X_h} = (X_{h1}, X_{h2}, \ldots\ldots, X_h{}_{n/2})^T$ and maximum possible score vector for *g*-th and *h*-th sub-test is $\boldsymbol{I} = (\frac{n}{2}, \frac{n}{2}, \ldots\ldots, \frac{n}{2})^T$ since maximum possible score which can be obtained by an individual in a sub-test consisting of $\frac{n}{2}$ number of items is $\frac{n}{2}$.

Let θ_{Xg} be the angle between $\boldsymbol{X_g}$ and $\boldsymbol{I}$. Then, $Cos\theta_{Xg} = \frac{\sum X_{ig}}{\|X_g\|\sqrt{N}}$. Since length of the vector $\boldsymbol{I}$ is $\|\boldsymbol{I}\| = \frac{n\sqrt{N}}{2}$, Similarly, $Cos\theta_{Xh} = \frac{\sum X_{ih}}{\|X_h\|\sqrt{N}}$ where θ_{Xh} is the angle between $\boldsymbol{X_h}$ and $\boldsymbol{I}$.

Now $\overline{X_g} = \overline{X_h} \Longrightarrow \|X_g\| \, Cos\theta_{Xg} = \|X_h\| \, Cos\theta_{Xh}$ or $\frac{\|X_g\|}{\|X_h\|} = \frac{Cos\theta_{Xh}}{Cos\theta_{Xg}}$

Since parallel tests have equal mean and equal variance,

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

i.e. for parallel tests, $\|X_g\|^2 = \|X_h\|^2 \Longrightarrow Cos\,\theta_{Xg} = Cos\,\theta_{Xh} \Longrightarrow \theta_{Xg} = \theta_{Xh}$

Thus, two vectors representing parallel tests are of equal length and makes equal angle with the Max. Possible vector. The property may be used for testing whether two sub-tests are parallel in terms of equality of $Cos\,\theta_{Xg}$ and $Cos\,\theta_{Xh}$, i.e. cosine similarity. Spruill, (2007) has shown that under H_0 distribution of dot product of two independent random vectors, each with unit length is well approximated by the Normal distribution for large *N*.

The method based on Cosine similarity may not assume normal distribution of X_g and X_h and may offer a better solution to the problem of statistically assessing parallelism with prescribed accuracy.

4. Discussions and Conclusions:

The paper presented methods of computing sample estimates of S_E^2 , ${S_T}^2$ (even if true scores of individuals taking the test are not known) and test reliability as per the theoretical definition from a single administration in terms of length of score vectors of two parallel subtests and angle between such vectors. The method also helps to have unbiased and consistent estimates of σ_T^2 and σ_E^2 for the population. Computation of reliability as per theoretical definition helps to test reliability is equal to one which is equivalent to testing H_0: $\sigma_X^2 = \sigma_T^2$. Rejection of the hypothesis H_0: $\sigma_X^2 = \sigma_T^2$ indicates higher values of σ_E^2 and poor quality of the test, Reporting of SD of true score or SD of error score and theoretically defined reliability is recommended for a test.

Linear regression of *T* on *X* for estimating true score is found to have desirable properties. Using $\hat{T}$ of each individual taking the test, one may undertake computation of the probability that the percentile true score of the *i*-th examinee is *t*, given the observed percentile score of the examinee is *x* and reliability is *r*, i.e. Pr.($T \leq t | r_{tt} = r, X \leq x$). Confidence interval of true score using *SD* of residual (S_ϵ) and theoretical definition of test reliability, is likely to work better since S_ϵ^2 is less than the test error variance S_E^2. In addition, to predict an interval of future values of *T* for a given X_0, a prediction interval in terms of S_ϵ can be used. Prediction interval is designed to cover a "moving target", the random future value of T, while the confidence interval is designed to cover the "fixed target", the average (expected) value of *T* for a given X_0. However, length of both confidence and prediction intervals will get increased as $(X_0 - \bar{X})^2$ is increased. Confidence interval of test reliability was found in an innovative fashion using the fact the slope of regression equation of *T* on *X* is the test reliability as per theoretical definition.

Simultaneous testing of single multidimensional hypothesis of equality of mean, variance and correlation can also be carried out by testing equality of regression line of *X* on X_g and *X* on X_h by ANOVA or by testing equality of two correlations r_{XX_g} and r_{XX_h} or by testing normality of $Z = (X_g - X_h)$ or by Cosine similarity (without assuming normal distribution of X_g and X_h). Identification of items, deletion of which will improve parallality of two sub-tests and/or make the test reliability robust may be approached using MTS, by selecting suitably the Normal group.

While in this work, the classically defined reliability of a test with binary items (one right and rest wrong) was discussed, a future study could be to find bias of the classically defined reliability with simulated reliabilities and compare with other methods of obtaining reliability.

5. Declaration:

Acknowledgement: Nil

Funding details: No funds, grants, or other support was received

Conflict of interests: The author has no conflicts of interest to declare

Competing interests: The authors report there are no competing interests to declare

Ethical Statement: This is a methodological paper and no ethical approval is required

Availability of data and material: Nil

Code availability: No application of software package or custom code

Authors' contributions: Sole author contributing at each stage of preparation of the manuscript from Conceptualization, Methodology and Writing

References

Chakrabartty, S. N. (2021): Angular similarity in test parameters, Methodological Innovations, 14(1). https://doi.org/10.1177/2059799120987786.

Cho, E., (2016): Making Reliability Reliable: A Systematic Approach to Reliability Coefficients, Organizational Research Methods, 19(4), 651-682.

Cortina J. (1993): What is coefficient alpha: an examination of theory and applications. Journal of Applied Psychology.78:98-104. https://doi.org/10.1037/0021-9010.78.1.98

Cressie, Noel (1979): A quick and easy empirical Bayes estimate of True scores, Sankhya, Vol. 41, Series B, 101-108

Eisinga, R., Te Grotenhuis, M., Pelzer, B. (2013): The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown? International Journal of Public Health, 58(4); 637-642. https://doi.org/10.1007/s00038-012-0416-3.

Feldt, L. S., Steffan, M., & Gupta, N. C. (1985): A comparison of five methods for estimating the standard error of measurement at specific score levels. Applied Psychological Measurement, 9, 351- 361.

Graham JM, Liu YJ, Jeziorski JL.(2006): The dyadic adjustment scale: a reliability generalization meta-analysis. J Marriage Fam. 68(3):701–717. https://doi.org/10.1111/j.1741-3737.2006.00284.x

Green, S. & Thompson M. (2005): Structural equation modeling in clinical psychology research In: Roberts M, Ilardi S, editors. Handbook of research in clinical psychology.Wiley-Blackwell

Guttman, L. (1945): A basis for analyzing test-retest reliability. Psychometrika, 10, 255-282.

Hedge, C., Powell, G., & Sumner, P. (2018): The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behavior Research Methods, 50(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Jackson, P.H. (1973): The estimation of true score variance and error variance in the classical test theory model. Psychometrika 38, 183–201. https://doi.org/10.1007/BF02291113

Kline, T. (2005): Psychological testing: A practical approach to design and evaluation. Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781483385693

Kristof, Walter (1969): Estimation of True Score and Error Variance for Tests under Various Equivalence Assumptions. Psychometrika 34, 489–507 https://doi.org/10.1007/BF02290603

Leininger, T. (2013): Unit 6: Simple linear regression [Lecture 3: Confidence and prediction intervals for SLR] Duke University.

Lord, F. M. (1959): Tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 19, 233-239.

Lord, F.M. & Novick, M.R. (1968): Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Miguel, A. & Garcia, P (2013): Statistical criteria for parallel tests: A comparison of accuracy and power, Behavior Research Methods, 45 (4), 999 - 1010

Panayides, P. (2013): Coefficient Alpha Interpret With Caution, Europe's Journal of Psychology, 9(4); 687-696. https://doi.org/10.5964/ejop.v9i4.653

Rao, C. R. (1952): Advanced Statistical Methods in Biometric Research. John Wiley & Sons.

Ritter, N. (2010): Understanding a widely misunderstood statistic: Cronbach's alpha.

Paper presented at Southwestern Educational Research Association (SERA) Conference, USA, (ED526237).

Rodgers, Joseph Lee and Nicewander, W. Alan (1988): Thirteen Ways to Look at the Correlation Coefficient, The American Statistician, 42(91), 59-66

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

Schmitt N. (1996): Uses and abuses of coefficient alpha. Psychological Assessment. 8(4). 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Shavelson, R. J., and Webb, N. M. (2012): Generalizability Theory. In J. Green, G. Camilli, & P. Elmore (Eds.), Handbook of complementary methods in education research (3rd ed.). Routledge.

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

Spruill, Marcus 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

Ten Berge, J.M.F., Snijders, T.A.B. & Zegers, F.E. (1981): Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. Psychometrika, 46, 201-213.

Teo T and Fan X (2013): Coefficient alpha and beyond: Issues and alternatives for educational research. Asia Pacific Education Review 22: 209–213.

Trafimow, David (2014): Estimating true standard deviations, Frontiers of Psychology, Vol. 5, Article 235, https://doi.org/10.3389/fpsyg.2014.00235

Verhelst, N. D. (2000): Estimating the Reliability of a Test from a Single Test Administration, National Institute for Educational Measurement, The Netherlands

Webb, Noreen, Richard, Shavelson, and Haertel, Edward. (2006): 4 Reliability Coefficients and Generalizability Theory. Handbook of Statistics. 26. 81-124. https://doi.org/10.1016/S0169-7161(06)26004-8.

Weisstein, Eric W. (2003): Sample Variance Distribution. MathWorld--A Wolfram Web Resource. https://mathworld.wolfram.com/SampleVarianceDistribution.html

Williams DR, Martin SR, Rast P. (2022): Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models. Behav Res Methods. 54(3):1272-1290. https://doi.org/10.3758/s13428-021-01646-x.

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