Reconceptualization of Coefficient Alpha Reliability for Test Summed and Scaled Scores

Abstract

Coefficient alpha reliability persists as the most common reliability coefficient reported in research. The assumptions for its use are, however, incomprehensively well-understood. The current paper challenges the commonly used expressions of coefficient alpha and argues that while these expressions are correct when estimating reliability for summed scores, they are not appropriate to extend coefficient alpha to correctly estimate the reliability for nonlinearly-transformed scaled scores such as percentile ranks and stanines. The current paper reconceptualizes coefficient alpha as a complement of the ratio of two unbiased estimates of the summed score variance. These include conditional summed score variance assuming uncorrelated item scores (gives the error score variance) and unconditional summed score variance incorporating intercorrelated item scores (gives the observed score variance). Using this reconceptualization, a new equation of coefficient generalized alpha is introduced for scaled scores. Two applications (cognitive and psychological assessments) are used to compare the performance (estimation and bootstrap confidence interval) of the reliability coefficients for different scaled scores. Results support the new equation of coefficient generalized alpha and compare it to coefficient generalized beta for parallel test forms. Coefficient generalized alpha produced different reliability values which were larger than coefficient generalized beta for different scaled scores.
Keywords: reliability, coefficient alpha, coefficient generalized alpha, coefficient beta, coefficient generalized beta, summed scores, scaled scores.

In different educational, psychological, social, medical and other fields, raw scores (summed scores) on tests and measurements are typically transformed to scaled scores for appropriate interpretations and decision-making purposes. Currently, almost all large-scale tests use scaled scores in their score reports. Examples of scaled scores include percentile ranks, age equivalents, standardized scores, and normalized scores. Reliability of the aforementioned scores was investigated abundantly in previous research studies (e.g., Almehrizi, 2013, 2016; Brennan & Lee, 1999; Feldt & Qualls, 1998; Kolen, et. al., 1992; Kolen & Lee., 2011; Kolen, et. al., 2012; Lee, 2007) which concluded that there is a need for investigating the reliability of all types of test scaled scores besides the reliability of summed scores. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council of

Measurement in Education, 2014) emphasize that reliability should be examined for all scores that are employed in test reports for better interpretation and utilization of test scores.

Reliability of summed scores do not describe the reliability of scaled scores except for those scores that result from linear transformations of summed scores. Test-retest reliability of summed scores, for example, is the correlation between the two summed scores obtained from two independent administrations of test items to the same group of examinees. Test-retest reliability of scaled scores is the correlation between the two scaled scores following the transformation of the summed scores obtained from two independent administrations of test items to the same group of examinees. These test-retest reliabilities are not equal for nonlinear scaled scores. Nonlinear scaled scores including percentile ranks, age equivalents, and normalized scores could indicate larger or smaller reliability values than summed scores (Almehrizi, 2013; Kolen, et. al., 1992; Kolen et. al., 1996).

As coefficient alpha (Cronbach, 1951) is commonly used to estimate the reliability of test scores, Almehrizi (2013) presented a generalization of coefficient alpha (coefficient generalized alpha) to estimate the internal consistency reliability for scaled scores under the same assumptions as coefficient alpha for summed scores. Both coefficient alpha for summed scores and coefficient generalized alpha for scaled scores assume that the test forms are essentially tau-equivalent in which all items measure a common construct, have equal true scores, and have unequal error scores.

In a matrix of responses, X_{pi}, for N examinees ($p = 1,2,\dots,N$) on K items ($i = 1,2,\dots,K$); all scored with J score points ($j = 1,2,\dots,J$), the three common expressions of coefficient alpha (Cronbach, 1951) for summed scores, X, are

$$\hat{\alpha} = \frac{K}{K-1}\left[1 - \frac{\sum_i \hat{\sigma}_i^2}{\hat{\sigma}^2(X)}\right] = 1 - \frac{\frac{K}{K-1}\left[\sum_i \hat{\sigma}_i^2 - \frac{1}{K}\hat{\sigma}^2(X)\right]}{\hat{\sigma}^2(X)} = \frac{\frac{K}{K-1}\left[\hat{\sigma}^2(X) - \sum_i \hat{\sigma}_i^2\right]}{\hat{\sigma}^2(X)}, \quad (1)$$

where $\sum_i \hat{\sigma}_i^2 = \frac{\sum_i \sum_p (X_{pi} - \bar{X}_i)^2}{N-1}$, and $\hat{\sigma}^2(X) = K^2 \frac{\sum_p (\bar{X}_p - \bar{X})^2}{N-1}$. Here, $\bar{X}_i$ is the mean score for an item, i, $\bar{X}_p$ is the mean for an examinee p, and $\bar{X}$ is the grand mean for the matrix of X_{pi}. If the summed scores, X, are transformed to scaled scores, S, the three expressions of coefficient generalized alpha (Almehrizi, 2013) for scaled scores are,

$$\widehat{G\alpha} = \frac{K}{K-1}\left[1 - \frac{\epsilon_i^2(S)}{\hat{\sigma}^2(S)}\right] = 1 - \frac{\frac{K}{K-1}\left[\epsilon_i^2(S) - \frac{1}{k}\hat{\sigma}^2(S)\right]}{\hat{\sigma}^2(S)} = \frac{\frac{K}{K-1}\left[\hat{\sigma}^2(S) - \epsilon_i^2(S)\right]}{\hat{\sigma}^2(S)}, \quad (2)$$

where $\epsilon_i^2\ (S) = \frac{N}{N-1}[\sum_X S^2 f(X|\boldsymbol{\pi}_i) - [\sum_X S f(X|\boldsymbol{\pi}_i)]^2]$ and

$$\hat{\sigma}^2(S) = \frac{N}{N-1}[\sum_X S^2 f(X) - [\sum_X S f(X)]^2].$$

Here, $f(X|\boldsymbol{\pi}_i)$ is the conditional probability mass function (under the assumption of uncorrelated item scores) for the summed scores based on the proportions of all score points on each individual item, $\boldsymbol{\pi}_i$. $f(X)$ is the unconditional observed probability mass function for the summed scores in the sample.

Although coefficient generalized alpha gives different reliability estimates for different test scaled scores that are dissimilar from the reliability estimate of test summed scores, it was found to produce values exceeding the value of 1 for some scaled scores as discussed by Almehrizi (2013). This is unacceptable, however, in the reliability context. Such values occur when error score variances show negative estimate values. Almehrizi (2013) suggested setting such estimates of error variance at a value of 0, because the error score variance in the population cannot take negative values while the negative estimates could result from either a small or nonrandom sampling.

This rectification of the negative estimates of error score variance is an ad-hoc and remains unacceptable given the fact that the error score variance should not be negative by

definition even in sample estimation. The current paper explains the cause of the negative estimate of this error score variance in coefficient generalized alpha. Additionally, the paper presents a different reconceptualization of the error score variance associated with coefficient alpha and presents a new corrected equation for coefficient generalized alpha for scaled scores under the classical definition of reliability within the classical test theory framework. The relationship between reliability estimations under the essentially tau-equivalence assumption and the classical parallelism assumption is discussed for test scaled scores. Finally, two applications (cognitive assessment and psychological assessment) are used to compare the performance (estimation and bootstrap confidence interval) of the reliability coefficients for different scaled scores.

Reconceptualization of Coefficient Alpha for Summed Scores

Coefficient alpha continues to be incomprehensively understood (Barbera et. al., 2021; Cho & Kim, 2015; Green & Yang, 2009; Sijtsma, 2009) even though it is widely reported in research studies (Sijtsma & Pfadt, 2021). Barbera et. al. (2021) argued that it is a result of the ambiguous and imprecise language used to describe what information is communicated by an estimated value of alpha. Coefficient alpha has specific definitions of both true scores and error scores in the reliability context. Ellis (2021) reported that coefficient alpha can be differently interpreted using the definition of true scores in three test theories: Classical test theory, generalizability theory and latent trait theory.

From the classical definition of reliability in the classical test theory ($r_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$), the error score variance in coefficient alpha for summed scores is defined as

$$\hat{\sigma}_\alpha^2(E) = \hat{\sigma}^2(X)(1-\hat{\alpha}) = \frac{K}{K-1}\left[\sum_i \hat{\sigma}_i^2 - \frac{1}{K}\hat{\sigma}^2(X)\right]. \quad (3)$$

$\hat{\sigma}_\alpha^2(E)$ consists of two components: the sum of item variances, $\sum_i \hat{\sigma}_i^2$, and the summed score variance, $\hat{\sigma}^2(X)$. The first component gives the summed score variance for uncorrelated item scores, whereas the second component provides the summed score variance for correlated item scores. Using a component that incorporates correlated item scores in estimating error score variance contradicts the independence assumption of item error scores in classical test theory, which results in a paradox.

As a special case, this form gives an accurate estimate of the error score variance for summed scores only. This is algebraically true because the grand mean of summed scores ($\bar{X}$) is equal to the average of person means ($\bar{X}_p$) and to the average of item means ($\bar{X}_i$). However, because scaled scores are transformations of summed scores and not transformations of item scores (an exception is only for linear scaled scores), this relationship between grand mean and person and item means is not applicable. This form of the error score variance for nonlinearly-transformed scaled scores, if used, produces an estimate of the error score variance that incorporates correlation among items and contradicts the assumption of independent error scores. The error score variance is either underestimated or overestimated and could have negative estimates for some scaled scores. This would result in a biased coefficient generalized alpha and could exceed a value of 1.

$\hat{\sigma}_\alpha^2(E)$ can be written in a different way that is congruent with the classical definition of independent error scores,

$$\hat{\sigma}_\alpha^2(E) = \frac{KN}{N-1}\left[\frac{1}{N}\sum_p \frac{\sum_i (X_{pi}-\bar{X}_p)^2}{K-1} - \frac{\sum_i (\bar{X}_i-\bar{X})^2}{K-1}\right]. \quad (4)$$

This form consists of two terms: the average of within person variances and the item mean variance. Moreover, $\hat{\sigma}_\alpha^2(E)$ can be written in more useful form,

$$\hat{\sigma}_{\alpha}^{2}(E) = \frac{N}{N-1}\frac{K}{K-1}\left[\frac{\sum_p \sum_i (X_{pi}-\bar{X}_p)^2}{N} + \frac{\sum_i \sum_p (X_{pi}-\bar{X}_i)^2}{N} - \frac{\sum_i \sum_p \left(X_{pi}-\bar{X}\right)^2}{N}\right], \quad (5)$$

because $\sum_i \sum_p \left(X_{pi} - \bar{X}\right)^2 = \sum_i \sum_p \left(X_{pi} - \bar{X}_i\right)^2 + \sum_i \sum_p (\bar{X}_i - \bar{X})^2$.

Examining the three components in equation (5) shows three properties: 1) all components incorporate uncorrelated item scores and conform to the assumption of independent error scores for classical test theory, 2) all components can be estimated for nonlinear scaled scores, 3) all components are within-person variances using different mean scores (person mean, item mean, grand mean).

Using the item by person data matrix, the first term is the average of within-person variances using person means, the second is the sum of item variances (or the average of within-person using item means), and the third is a function of total variance (or the average of within-person using grand mean). These three terms estimate conditional summed score variance under the assumption of uncorrelated item scores using different conditioning functions of summed scores. That is,

$$\epsilon_p^2(X) = \frac{\sum_p \sum_i (X_{pi}-\bar{X}_p)^2}{N} = \frac{1}{N}\sum_p \left[\sum_X X^2 f(X|\boldsymbol{\pi}_p) - \left[\sum_X X f(X|\boldsymbol{\pi}_p)\right]^2\right], \quad (6)$$

$$\epsilon_i^2(X) = \frac{\sum_i \sum_p (X_{pi}-\bar{X}_i)^2}{N} = \sum_X X^2 f(X|\boldsymbol{\pi}_i) - \left[\sum_X X f(X|\boldsymbol{\pi}_i)\right]^2, \quad (7)$$

$$\epsilon^2(X) = \frac{\sum_i \sum_p (X_{pi}-\bar{X})^2}{N} = \sum_X X^2 f(X|\boldsymbol{\pi}) - \left[\sum_X X f(X|\boldsymbol{\pi})\right]^2 . \quad (8)$$

Hence, $\hat{\sigma}_{\alpha}^{2}(E_x) = \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(X) + \epsilon_i^2(X) - \epsilon^2(X)\right]$.

Here, $f(X|\boldsymbol{\pi}_p)$, $f(X|\boldsymbol{\pi}_i)$, and $f(X|\boldsymbol{\pi})$ represent the conditional probability mass functions for the summed scores on the proportions of all score points for three item responses: (1) person response patterns, (2) item response patterns, and (3) total response patterns. These three terms $(\epsilon_p^2(X), \epsilon_i^2(X), \epsilon^2(X))$ provide the estimate of the conditional summed score variance under the assumption of independent item error scores. Note that $f(X|\boldsymbol{\pi}_p)$

uses the proportions of all score points in the observed response pattern for each individual p, $f(X|\boldsymbol{\pi}_i)$ employs the proportions of all score points in the observed response pattern for each item i, and $f(X|\boldsymbol{\pi})$ utilizes the proportions of all score points in the observed response pattern matrix.

Conversely, the unbiased observed summed score variance, $\hat{\sigma}^2(X)$, is the unconditional summed score variance that incorporates the intercorrelation among item scores. That is

$$\hat{\sigma}^2(X) = \frac{N}{N-1}\sigma^2(X) = \frac{N}{N-1}[\sum_X X^2 f(X) - [\sum_X X f(X)]^2], \quad (9)$$

where $f(X)$ is the observed probability mass function for summed scores in the sample. Hence, using the classical definition of reliability in classical test theory, the error score variance and the true score variance for summed scores associated with coefficient alpha are,

$$\hat{\sigma}_\alpha^2(E) = \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(X) + \epsilon_i^2(X) - \epsilon^2(X)\right],$$

$$\hat{\sigma}_\alpha^2(T) = \hat{\sigma}^2(X) - \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(X) + \epsilon_i^2(X) - \epsilon^2(X)\right]$$

and coefficient alpha for summed scores is,

$$\hat{\alpha} = 1 - \frac{\frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(X)+\epsilon_i^2(X)-\epsilon^2(X)\right]}{\hat{\sigma}^2(X)} = \frac{\hat{\sigma}^2(X)-\frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(X)+\epsilon_i^2(X)-\epsilon^2(X)\right]}{\hat{\sigma}^2(X)}. \quad (10)$$

This form suggests that need for a new conceptualization of coefficient alpha for summed scores. It is the complement of the ratio of two unbiased estimators of summed score variances conditional summed score variances which assume uncorrelated item scores, $\hat{\sigma}_\alpha^2(E)$, and the unconditional summed score variance incorporating intercorrelated item scores, $\hat{\sigma}^2(X)$.

Reconceptualization of Coefficient Alpha for Scaled Scores

The new conceptualization of coefficient alpha facilitates the estimation of coefficient alpha reliability for scaled scores. All terms in equation (10) are about the summed scores not item scores. The transformation function used to obtain scaled scores for summed scores can be utilized to obtain the corresponding terms for scaled scores by substituting each summed score by its scaled score in equations (6) to (9). That is,

$$\epsilon_p^2(S) = \frac{1}{N}\sum_p \left[\sum_X S^2 f(X|\boldsymbol{\pi}_p) - \left[\sum_X S f(X|\boldsymbol{\pi}_p)\right]^2\right], \quad (11)$$

$$\epsilon_i^2(S) = \sum_x S^2 f(X|\boldsymbol{\pi}_i) - [\sum_x S f(X|\boldsymbol{\pi}_i)]^2 , \quad (12)$$

$$\epsilon^2(S) = \sum_x S^2 f(X|\boldsymbol{\pi}) - [\sum_x S f(X|\boldsymbol{\pi})]^2 , \quad (13)$$

$$\hat{\sigma}^2(S) = \frac{N}{N-1}[\sum_x S^2 f(X) - [\sum_x S f(X)]^2]. \quad (14)$$

$\epsilon_p^2(S)$, $\epsilon_i^2(S)$, and $\epsilon^2(S)$ are the scaled score variances assuming uncorrelated item scores and conditioned on the proportions of all score points for three response patterns: person response patterns, item response patterns, and total response patterns, respectively. $\hat{\sigma}^2(S)$ is the unconditional scaled score variance that incorporates the intercorrelations among the item scores. All these four scaled score variances are larger than zero.

Hence, using the classical definition of reliability in the classical test theory, the error score variance and the true score variance for scaled scores are,

$$\hat{\sigma}_{G\alpha}^2(E) = \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(S) + \epsilon_i^2(S) - \epsilon^2(S)\right],$$

$$\hat{\sigma}_{G\alpha}^2(T) = \hat{\sigma}^2(S) - \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(S) + \epsilon_i^2(S) - \epsilon^2(S)\right]$$

and coefficient generalized alpha for scaled scores is,

$$\widehat{G\alpha} = 1 - \frac{\frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(S)+\epsilon_i^2(S)-\epsilon^2(S)\right]}{\hat{\sigma}^2(S)} = \frac{\hat{\sigma}^2(S) - \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon_p^2(S)+\epsilon_i^2(S)-\epsilon^2(S)\right]}{\hat{\sigma}^2(S)}. \quad (15)$$

Similarly, coefficient generalized alpha for scaled scores is defined as the complement of the ratio of two scaled score variances: conditional scaled score variances assuming

uncorrelated item scores, $\hat{\sigma}^2_{G\alpha}(E)$, and unconditional scaled score variance incorporating intercorrelated item scores, $\hat{\sigma}^2(S)$.

Since summed scores are a special case of scaled scores (identity transformation of summed scores), coefficient generalized reliability if applied to summed scores gives exact estimates as coefficient alpha. Hypothetically, $\hat{\sigma}^2_{G\alpha}(E)$ should range between zero (hence, $\widehat{G\alpha}$ =1) and $\hat{\sigma}^2(S)$ (hence, $\widehat{G\alpha}$ =0). However, $\widehat{G\alpha}$ could have negative values indicating that $\hat{\sigma}^2_{G\alpha}(E)$ could exceed $\hat{\sigma}^2(S)$. This happens when test items show negative intercorrelations. If items are consistent, they should measure one construct and thus items display no negative intercorrelations.

Estimation of Conditional Probability Mass Functions

The three conditional summed score variances (and scaled score variances) require estimating three conditional probability mass functions for summed scores under the assumption of uncorrelated item scores using the proportions of all score points for three response patterns: person response patterns, item response patterns, and total response patterns. All of these conditional probability mass functions for summed scores can be obtained using the same recursion formula adapted by Almehrizi (2013) with different inputs of the proportions of item score points [say P_{ij} for $c_1, c_2, \dots, c_J$, where J is the number of score points]. There are three conditioning matrices of proportions of item score points for the three response patterns as follows: $P_{ij} = \pi_{pj}$ for $f(X|\boldsymbol{\pi}_p)$ when estimating $\epsilon^2_p(X)$ and $\epsilon^2_p(S)$ for each single examinee, $P_{ij} = \pi_{ij}$ for $f(X|\boldsymbol{\pi}_i)$ when estimating $\epsilon^2_i(X)$ and $\epsilon^2_i(S)$, and $P_{ij} = \pi_j$ for $f(X|\boldsymbol{\pi})$ when estimating $\epsilon^2(X)$ and $\epsilon^2(S)$.

Let $f(X|\boldsymbol{P_{ij}})$ be the desired conditional probability mass functions for summed scores, X, on the conditioning matrix of proportions of item score points, $\boldsymbol{P_{ij}}$. To obtain $f(X|\boldsymbol{P_{ij}})$ through the recursion formula, define X_i as a random variable of raw scores on the

first i items on the test (X_i ranges between ic_1 and ic_J). Now, let $f(X_i|\boldsymbol{P_{ij}})$ represent the probability mass function of summed scores, X_i, on a test of i items. For a test of one item, i = 1, entered into the formula,

$$f(X_1|\boldsymbol{P_{ij}}) = P_{1j}, \quad \text{for} \quad X_1 = c_1, c_2 \ \ldots c_J. \tag{16}$$

For the next $i > 1$, the recursion formula is as follows:

$$f(X_i|\boldsymbol{P_{ij}}) = \sum_j f(X_{i-1} = X_i - c_j|\boldsymbol{P_{ij}})P_{ij}, \text{ for } X_i = ic_1, ic_1 + d \ \ldots ic_J. \tag{17}$$

To use this recursion formula, items are entered into the recursion formula in any order, beginning with any item; and the formula can be applied recurrently by increasing i on each recurrence until entering all K items. The process ends after $i = K$, which gives the required $f(X|\boldsymbol{P_{ij}})$. That is, $f(X|\boldsymbol{P_{ij}}) = f(X_K|\boldsymbol{P_{ij}})$.

It is noteworthy to mention that the same recursion formula can be used to estimate the three conditional probability mass functions for summed scores but with using different conditioning matrix of the proportions of item score points, P_{ij}. The recursion formula is applied one time to obtain $f(X|\boldsymbol{\pi})$ using $P_{ij} = \pi_j$ for all items (equal for all items), one time to obtain $f(X|\boldsymbol{\pi}_i)$ using $P_{ij} = \pi_{ij}$ for each item (different by items), and one time for obtaining $f(X|\boldsymbol{\pi}_p)$ for each individual p, using $P_{ij} = \pi_{pj}$ for all items (equal for all items). There are N times of recursion formula to obtain $f(X|\boldsymbol{\pi}_p)$;each for a single examinee.

Relationship with Coefficient Beta for Summed and Scaled Scores

Almehrizi (2021) presented coefficient beta for summed scores and scaled scores on a test of any number of scoring patterns as an extension of KR-21 reliability for Kuder and Richardson (1937). This is an appropriate estimate of reliability of summed scores and scaled scores when the test forms are classically-parallel equivalent where test items measure the same construct and have equal true scores and equal uncorrelated error

scores. Using the classical definition of reliability in the classical test theory, the error score variance and the true score variance for summed scores are,

$$\hat{\sigma}_{\beta}^{2}(E) = \frac{K}{K-1}\epsilon_{p}^{2}(X) \text{ and } \hat{\sigma}_{\beta}^{2}(T) = \hat{\sigma}^{2}(X) - \frac{K}{K-1}\epsilon_{p}^{2}(X), \tag{18}$$

and coefficient beta for summed scores is,

$$\hat{\beta} = 1 - \frac{\frac{K}{K-1}\epsilon_{p}^{2}(X)}{\hat{\sigma}^{2}(X)} = \frac{\hat{\sigma}^{2}(X) - \frac{K}{K-1}\epsilon_{p}^{2}(X)}{\hat{\sigma}^{2}(X)}. \tag{19}$$

For scaled scores, the error score variance and the true score variance are,

$$\hat{\sigma}_{G\beta}^{2}(E) = \frac{K}{K-1}\epsilon_{p}^{2}(S) \text{ and } \hat{\sigma}_{G\beta}^{2}(T) = \hat{\sigma}^{2}(S) - \frac{K}{K-1}\epsilon_{p}^{2}(S), \tag{20}$$

and coefficient generalized beta for scaled scores is,

$$\widehat{G\beta} = 1 - \frac{\frac{K}{K-1}\epsilon_{p}^{2}(X)}{\hat{\sigma}^{2}(S)} = \frac{\hat{\sigma}^{2}(S) - \frac{K}{K-1}\epsilon_{p}^{2}(X)}{\hat{\sigma}^{2}(S)}. \tag{21}$$

From equations (19) and (21), both coefficient beta and coefficient generalized beta are interpreted as the complement of the ratio of two summed/scaled score variances: conditional summed/scaled score variance assuming uncorrelated item scores and unconditional summed/scaled score variance incorporating intercorrelated item scores.

When comparing coefficient alpha and coefficient beta for summed scores, the two coefficients use different estimators of both error score variance and true score variance. Coefficient beta is usually smaller than or equal to coefficient alpha such as,

$$\hat{\beta} = \hat{\alpha} - \frac{\frac{N}{N-1}\frac{K}{K-1}\left[\epsilon^{2}(X) - \epsilon_{i}^{2}(X) - \frac{1}{N}\epsilon_{p}^{2}(X)\right]}{\hat{\sigma}^{2}(X)}. \tag{22}$$

Additionally, the difference between coefficient generalized alpha and coefficient generalized beta for scaled scores is a function of the three conditional scaled score variance and the unconditional scaled score variance. Similarly, coefficient generalized beta for scaled scores is usually smaller than or equal to coefficient generalized alpha such as,

$$\widehat{G\beta} = \widehat{G\alpha} - \frac{\frac{N}{N-1}\frac{K}{K-1}\left[\epsilon^{2}(S) - \epsilon_{i}^{2}(S) - \frac{1}{N}\epsilon_{p}^{2}(S)\right]}{\hat{\sigma}^{2}(S)}. \tag{23}$$

This relationship between coefficient alpha and coefficient beta for both summed scores and scaled scores conforms to the well-established fact in the classical test theory. The error score variance when the test forms are essentially tau equivalent is smaller than the error score variance when the test forms are parallel equivalent (Feldt, 1984; Feldt & Qualls, 1998). That is,

$$\hat{\sigma}_{\alpha}^{2}(E) = \hat{\sigma}_{\beta}^{2}(E) - \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon^{2}(X) - \epsilon_{i}^{2}(X) - \frac{1}{N}\epsilon_{p}^{2}(X)\right], \quad (24)$$

$$\hat{\sigma}_{G\alpha}^{2}(E) = \hat{\sigma}_{G\beta}^{2}(E) - \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon^{2}(S) - \epsilon_{i}^{2}(S) - \frac{1}{N}\epsilon_{p}^{2}(S)\right]. \quad (25)$$

Also, the two coefficients do not only employ different estimators of error score variance but also different estimators of true score variance. The true score variance when the test forms are essentially tau equivalent is larger than the true score variance when the test forms are parallel equivalent by an amount related to differences of item averages. That is

$$\hat{\sigma}_{\alpha}^{2}(T) = \hat{\sigma}_{\beta}^{2}(T) + \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon^{2}(X) - \epsilon_{i}^{2}(X) - \frac{1}{N}\epsilon_{p}^{2}(X)\right], \quad (26)$$

$$\hat{\sigma}_{G\alpha}^{2}(T) = \hat{\sigma}_{G\beta}^{2}(T) + \frac{N}{N-1}\frac{K}{K-1}\left[\epsilon^{2}(X) - \epsilon_{i}^{2}(S) - \frac{1}{N}\epsilon_{p}^{2}(S)\right]. \quad (27)$$

This amount of difference between the two error score variances is identical to the difference between the two true score variances. This is correct for both summed scores and scaled scores.

Application 1

Measures and Participants

The new formula of coefficient generalized alpha for summed scores and scaled scores was applied to numerical ability subtest on the cognitive abilities assessment using R programming language. The cognitive abilities assessment aims to assess students from Kindergarten to 6th grade in three cognitive abilities: Verbal, Numerical and Spatial (Al-zayat & Almehrizi, 2011). Each ability test has 30 multiple-choice items that are scored

dichotomously with 0 (incorrect answer) and 1 (correct answer). The summed scores range between 0 and 30 where higher scores indicate higher ability level displayed by the child. Four types of scaled scores were reported: PR scores (percentile ranks), RSS scores (rounded standard scores with a mean of 100 and standard deviation of 10), NDS scores (rounded normalized developmental scores with a mean of 200 and standard deviation of 15), NAL scores (five numerical ability levels: very low ability (0-6→1), low ability (7-10→2), medium ability (11-22→3), high ability (23-27→4), and very high ability (28-30→5)). The collected data were for 4206 students from 1st and 2nd grade from a representative sample as part of a standardization study of the cognitive abilities assessment in Oman. The sample consisted of 1998 males and 2208 females.

Results

Table 1 presents means, standard deviations, coefficient generalized alpha and coefficient generalized beta, and the square root of the error score variance (standard errors of measurement) for summed and scaled scores of numerical ability scores. For summed scores and all scaled scores, the estimates resulted from the new equation of coefficient generalized alpha were larger than the estimated produced by coefficient generalized beta. Results showed that different scales had different reliability estimates compared to the summed scores. The reliability estimates scores by the new equation of coefficient generalized alpha and coefficient generalized beta for different scaled scores had the following descending ordering: PR, summed, RDS, RSS, and NAL scores.

Table 1

Descriptive statistics, coefficient generalized alpha and beta and SEM for summed and scaled scores for numerical ability

Statistics	Summed	RSS	PR	RDS	NAL
M	15.6510	99.0221	46.6153	202.7953	2.9318
SD	5.8560	8.4427	25.0530	15.4383	0.7436
Old $\widehat{G\alpha}$	0.8519	0.9149	0.8420	0.8787	0.9985
New $\widehat{G\alpha}$	0.8519	0.7869	0.8531	0.8163	0.6588
$\widehat{G\beta}$	0.8089	0.7533	0.8057	0.7754	0.6293
Old $\hat{\sigma}_{G\alpha}(E)$	2.2536	2.4629	9.9584	5.3769	0.0288
New $\hat{\sigma}_{G\alpha}(E)$	2.2536	3.8974	9.6022	6.6169	0.4344
$\hat{\sigma}_{G\beta}(E)$	2.5599	4.1934	11.0432	7.3165	0.4527

The old and the new equations for coefficient generalized alpha for summed scores produced equal estimates of reliability (0.8519). For other scaled scores, these two equations of the coefficient generalized alpha produced different estimates of reliability. The old equation produced estimates that ranged between 0.8420 and 0.9985, whereas the new equation produced estimates that ranged between 0.6588 and 0.8531. The new equation gave a higher reliability value (0.8531) for PR scores than the old equation (0.8420) and a lower value (0.6293) for NAL scores than the old equation (0.9985).

The estimates of the standard errors of measurement showed similar patterns as their corresponding estimates of reliability coefficients. The standard errors of measurement associated with the new equation for coefficient generalized alpha were always smaller compared to coefficient generalized beta for all scaled scores.

Table 2 presents non-parametric bootstrap (100 replications) averages, standard errors, and 95% confidence intervals for coefficient generalized alpha and beta for summed and scaled scores for numerical ability subtest. The standard error of estimates for coefficient generalized alpha and beta were small for summed and all scaled scores. The lower and higher bounds for the 95% bootstrap confidence interval showed acceptable

performance of coefficient generalized alpha and beta for the summed and all scaled scores. As expected, the higher bound for the old equation of coefficient generalized alpha for NAL scores was larger than 1, which supported the previous discussed problem with the old equation.

Results also showed that there were no overlaps between the confidence intervals for the new coefficient generalized alpha and coefficient generalized beta for summed scores and all scaled scores except for NAL scores. This could be interpreted that coefficient generalized alpha values were significantly larger than coefficient generalized beta values indicating that numerical ability test conformed with the assumptions of tau-equivalent forms more than the assumptions of classically parallel forms.

Table 2

Non-parametric bootstrap averages, SE, and 95% confidence intervals for coefficient generalized alpha and beta for summed and scaled scores for Numerical Ability

	Statistics	Summed	T-score	PR	NSS	NAL
Old $\widehat{G\alpha}$	Average	0.8517	0.9144	0.8422	0.8791	0.9982
	SE	0.0032	0.0031	0.0029	0.0031	0.0021
	LB	0.8411	0.9061	0.8344	0.8712	0.9904
	HB	0.8600	0.9225	0.8505	0.8863	1.0022
New $\widehat{G\alpha}$	Average	0.8517	0.7864	0.8534	0.8167	0.6572
	SE	0.0032	0.0048	0.0030	0.0037	0.0093
	LB	0.8411	0.7739	0.8451	0.8075	0.6314
	HB	0.8600	0.7997	0.8616	0.8251	0.6803
$\widehat{G\beta}$	Average	0.8087	0.7525	0.8059	0.7760	0.6285
	SE	0.0044	0.0060	0.0042	0.0049	0.0103
	LB	0.7954	0.7370	0.7938	0.7655	0.5985
	HB	0.8193	0.7701	0.8157	0.7879	0.6519

SE: bootstrap estimation standard error, LB & HB: Lower bound and higher bound for 95% bootstrap CI

Application 2

Measures and Participants

The second data set was the reasoning subtest of the Learning Disabilities Diagnostic Inventory (LDDI). The tool was developed by Don Hammill and Brian Bryant in 1998

(Hammill and Bryant, 1998) and was standardized on a national sample of Omani students in Oman in 2019 (El-Keshky & Emam, 2015; Emam et. al., 2021). LDDI is a teacher/clinician rating scale that seeks to systematically identify specific learning disabilities based on a child's intrinsic processing difficulties by observing the child's day-to-day academic behaviours.. It has six independent subtests; each contains 15 items focusing on the neuropsychological aspects of specific learning disabilities as an intrinsic disorder. Teachers rate their students on each item using nine score points (1 to 9) reflecting the incidence rates of each behavior (from never to always). The summed scores for the intrinsic disorder score range between 15 and 135 where higher scores indicate higher disorder.

Data were collected for 413 students from 1st and 2nd grades from all over the country as part of a standardization study of the LDDI in Oman. The sample consisted of 200 males and 213 females. In addition to the summed scores, four scaled scores were used for interpretation purposes. They were PR scores (percentile ranks), T-scores (rounded standard scores with a mean of 50 and standard deviation of 10), stanine scores (scores between 1 and 9), and reasoning diagnostic category (RDC) scores (three-band categorization: Normal (15-29$\rightarrow$1), Borderline (30$\rightarrow$104), Abnormal (105-135$\rightarrow$3)).

Results

Table 3 displays means, standard deviations, coefficient generalized alpha and generalized beta, and the square root of error score variance for summed and scaled scores of the reasoning subscale. For summed scores and all scaled scores, the estimates resulting from the new equation of coefficient generalized alpha were larger than the estimates produced by the coefficient generalized beta. Results showed that different scaled scores had different reliability estimates than the summed scores. The reliability estimate scores by the new equation of coefficient generalized alpha and

coefficient generalized beta for different scales are in descending order as follows: summed, T-scores, PR, stanine, and RDC scores.

Coefficient generalized alpha for summed scores was 0.9831. However, the two equations of coefficient generalized alpha for all scaled scores produced different estimates of reliability. The old equation produced estimates that ranged between 0.9555 and 1.0678; whereas the new equation produced estimates that ranged between 0.9187 and 0.9824. The old equation gave an estimate of the coefficient generalized alpha for RDC scores exceeding 1.

The estimates resulting from the new equation of generalized alpha for summed scores and all scaled scores were larger than the estimates produced by coefficient generalized beta. However, the old equation of generalized alpha showed mixed results. They were larger than coefficient generalized beta for some scaled scores (e.g., summed, T-scores, RDC); but they were smaller for other scores (e.g., PR, Stanine). The standard errors of measurement associated with the coefficient generalized alpha were smaller than the coefficient generalized beta and showed similar patterns across different scaled scores.

Table 3

Descriptive statistics, coefficient generalized alpha and beta and SEM for summed and scaled scores for LDDI Reasoning

Statistics	Summed	T-score	PR	Stanine	RDC
M	75.1162	49.9831	50.3487	5.0169	2.1695
SD	33.5623	10.0299	29.1908	2.6008	0.5445
Old $\widehat{G\alpha}$	0.9831	0.9817	0.9703	0.9555	1.0678
New $\widehat{G\alpha}$	0.9831	0.9824	0.9785	0.9675	0.9187
$\widehat{G\beta}$	0.9818	0.9811	0.9769	0.9660	0.9186
Old $\hat{\sigma}_{G\alpha}(E)$	4.3687	1.3555	5.0344	0.5489	-
New $\hat{\sigma}_{G\alpha}(E)$	4.3687	1.3324	4.2801	0.4685	0.1552
$\hat{\sigma}_{G\beta}(E)$	4.5295	1.3794	4.4354	0.4797	0.1553

- error variance is negative and there is no SEM.

Table 4 presents non-parametric bootstrap (100 replications) averages, standard errors, and 95% confidence intervals for coefficient generalized alpha and beta for summed and scaled scores for the LDDI Reasoning. Results showed that coefficient generalized alpha and beta were consistent for summed and all scaled scores because the standard errors were small. The 95% confidence intervals showed acceptable lower and higher bounds for all scaled scores. The old equation of coefficient generalized alpha for RDC scores had both lower and higher bounds exceeding 1.

Results also showed that the confidence intervals for the new coefficient generalized alpha and coefficient generalized beta had large overlaps for summed scores and all scaled scores. This could be interpreted that coefficient generalized alpha values were similar to coefficient generalized beta values indicating that the assumptions of classically parallel forms might describe the structure of the LDDI reasoning.

Table 4

Non-parametric bootstrap averages, SE, and 95% confidence intervals for coefficient generalized alpha and beta for summed and scaled scores for LDDI Reasoning

	Statistics	Summed	T-score	PR	Stanine	RDC
Old $\widehat{G\alpha}$	Average	0.9833	0.9818	0.9699	0.9549	1.0673
	SE	0.0014	0.0014	0.0057	0.0062	0.0026
	LB	0.9789	0.9783	0.9563	0.9389	1.0572
	HB	0.9861	0.9853	0.9805	0.9684	1.0706
New $\widehat{G\alpha}$	Average	0.9833	0.9824	0.9783	0.9674	0.9199
	SE	0.0014	0.0014	0.0018	0.0025	0.0110
	LB	0.9789	0.9790	0.9742	0.9619	0.8928
	HB	0.9861	0.9857	0.9818	0.9737	0.9437
$\widehat{G\beta}$	Average	0.9820	0.9810	0.9767	0.9658	0.9198
	SE	0.0016	0.0015	0.0020	0.0027	0.0109
	LB	0.9770	0.9775	0.9724	0.9600	0.8928
	HB	0.9850	0.9847	0.9805	0.9726	0.9437

SE: bootstrap estimation standard error, LB & HB: Lower bound and higher bound for 95% bootstrap CI

Conclusions and Implications

The paper examined the commonly used expressions of coefficient alpha with test summed scores and argued that although these expressions are correct when estimating reliability for summed scores, they are not sufficient to correctly extend coefficient alpha to estimate the reliability of nonlinearly-transformed scaled scores such as percentile ranks, normalized standard scores, and stanines. The paper argued that these common equations of coefficient alpha erroneously led to a wrong extension to coefficient generalized alpha in a paper published by Almehrizi (2013).

As an alternative, the paper treated coefficient alpha as a complement of the ratio of two unbiased estimates of summed score variance: conditional summed score variance assuming uncorrelated item scores (gives the error score variance) and unconditional summed score variance incorporating intercorrelated item scores (gives the observed score variance). This reconceptualization of coefficient alpha should facilitate a correct extension to estimate reliability of nonlinearly scaled scores. The new equation of coefficient generalized alpha is also the complement of the ratio of two unbiased estimates of the scaled score variance: conditional scaled score variance assuming uncorrelated item scores and unconditional scaled score variance incorporating intercorrelated item scores.

To estimate the reliability of scaled scores, the old equation of coefficient generalized alpha (equation 2) should not be used and, instead, should be replaced with the new equation of coefficient generalized alpha (equation 15). The two data sets in this paper showed that the old equation could have unacceptable values and could obtain values exceeding 1 for some scaled scores especially with a few scaled scores such as NAL scores (5 scores for numerical ability levels) and RDC scores (three scores of reasoning disability categories).

The reliability is not just a property of test items, but also a property of the scale being used for test scores when the transformation is nonlinear. Each nonlinear transformation of test scores produces different reliability coefficients, which were different from the reliability for summed scores. The reliability values for scaled scores could be either smaller or larger than for summed scores. The two applications presented earlier (cognitive assessment and psychological assessment) showed that all scaled scores had different reliability values and they were smaller than those for the summed scores except the reliability for PR scores in the numerical ability assessment. Scaled scores with a few levels such as NAL scores and RDC scores showed generally smaller values of coefficient generalized alpha and beta than other scaled scores with more scores such as percentile ranks and NSS. These results were recurrent when item response theory models were used (Kolen et. al., 1996; Kolen et. al., 2012).

The reliability values depend on the transformation function and the interaction between the transformation functions and item intercorrelations. Transformation of test scores makes them have different averages and variances for different transformation functions. The changes in the test score variance occur with both unconditional and conditional test score variances. When the transformation function is linear, the ratio of changes in the unconditional and the conditional test score variances is equal to the ratio of the two variances for the summed scores and, hence, the reliability coefficients of test summed score and test linearly-scaled scores become equal. However, when the transformation function is nonlinear, the changes in unconditional and conditional test score variances are not equal and, hence, the ratio of the two variances for the new scaled scores is different from the ratio of the two variances for the summed scores resulting on different reliability for the scaled scores compared to summed scores. Each type of scaled

scores will have a different ratio of these two test score variances and hence different reliability estimates.

The presented reconceptualization of coefficient generalized alpha is similar to coefficient beta reliability for summed scores and its generalization to scaled scores. Both coefficient beta reliability and its generalization are also the complement of the ratio of two unbiased estimates of the summed/scaled score variance: conditional summed/scaled score variance assuming uncorrelated item scores and unconditional summed/scaled score variance incorporating intercorrelated item scores. The difference between coefficient alpha and coefficient beta (and their generalization for scaled scores) is contingent upon the conditional summed/scaled score variances assuming uncorrelated item scores that give the error summed/scaled score variance. Coefficient beta uses the conditioning of proportions of item score points from response pattern for each examinee, whereas coefficient alpha uses three conditioning matrices of proportions of item score points. These are: 1) proportions of item score points from response pattern for each examinee, 2) proportions of item score points obtained from response patterns for each item, and 3) proportions of item score points from response patterns for the whole data matrix of examinees by items.

Coefficient alpha and coefficient beta (and their generalizations) have different true score variance and error score variance. The true score variance in coefficient beta is equal to the true score variance in coefficient alpha minus the item effect variance and the error score variance in coefficient beta is equal to the error score variance in coefficient alpha plus the item effect variance. As a result, coefficient beta is smaller than coefficient alpha (and their generalizations) by the ratio of the item effect variance and observed score variance. This result agrees with previous literature about the error score variance (such as Brennan, 2001; 2011; Barbera et. al., 2021; Raykov, 1998).

Both coefficient alpha and coefficient beta (and their generalizations) use the classical definition of reliability and both assume that all items assess a common construct, the error scores are uncorrelated and the true scores are independent from the error scores. Hence, with both coefficients, the observed score variance is equal to the sum of true score variance and error score variance.

Equation (24) through equation (27) reveals that coefficient alpha and coefficient beta (and their generalizations) have different assumptions about the definition of true scores and error scores. Coefficient beta (and coefficient generalized beta) assumes that items are not differentiated based on their variances (equal item variances) and not differentiated based on their averages (equal item averages). This denotes that coefficient beta follows classically parallel assumptions. So coefficient beta considers any differences in item variance and item averages are not part of true scores but sources of error scores. Hence, the error score variance in coefficient beta incorporates differences of both item averages and item variances.

On the other side, coefficient alpha (and coefficient generalized alpha) assumes that items are not differentiated based on their variances (equal item variances) but differentiated based on their averages (unequal item averages). This denotes that coefficient alpha follows essentially tau-equivalent assumptions. So coefficient alpha considers any differences in item averages are part of true scores and not a source of error scores, whereas it considers any differences in item variance a source of error scores and are not part of true scores. Therefore, the error score variance in coefficient alpha excludes differences in item averages while it incorporates differences of item variances.

This implies that coefficient alpha requires integrating item effect variance (the difference in item averages) in the operational definition of the construct assessed by the

test, whereas coefficient beta excludes it from the operational definition of the construct assessed by the test. Hence, practitioners should specify their operational definition of the measured construct by the test so they can determine which reliability coefficient to use.

Results of the two applications showed also that the estimates of coefficient alpha and coefficient beta in the numerical ability assessment were significantly different given that there were no overlaps between their bootstrap confidence intervals for summed scores and scaled scores. Therefore, it could be inferred that coefficient alpha gave a better estimate of the reliability of numerical ability test. For LDDI reasoning, the bootstrap confidence intervals for the estimates of coefficient alpha and coefficient beta overlapped for summed scores and scaled scores indicating that coefficient beta was a more suitable estimate of reliability coefficient.

Although the recursion formula outlined earlier can be used to obtain the conditional probability mass functions for summed scores for tests of items with either equal or different scoring points, both coefficients alpha and coefficient beta (and their generalizations) limit the application of this recursion formula to only those tests of items with similar item scoring points. For tests with mixed item formats (e.g., multiple choice items and essay items), the assumptions of coefficient alpha and beta are violated and, hence, they cannot be used to estimate reliability for summed scores or scaled scores in such tests.

Future research should investigate the use of the new reconceptualization of coefficient alpha and generalized alpha for different scaled scores with various assessment tools in different fields such as educational and achievement tests, medical assessments, social and economic studies, and engineering and industrial fields. There is a need for simulation studies to assess the statistical properties of coefficient generalized alpha and beta for

different scaled scores under different test conditions including test length, item scoring, sample size, score distribution, and conformity to assumption of test forms. In addition, the effect of estimating coefficient generalized alpha for different scaled scores should be investigated when its assumptions are violated including correlated error scores and dimensionality.

References

Almehrizi, R. S. (2013). Coefficient alpha and reliability of scaled scores. *Applied Psychological Measurement, 37*, 438-459. https://doi.org/10.1177/0146621613484983

Almehrizi, R. S. (2016). Normalization of mean squared differences to measure agreement for continuous data. *Statistical Methods in Medical Research, 25,* 1975-1990. https://doi.org/10.1177/0962280213507506

Almehrizi, R. S. (2021). Coefficient beta as extension of KR-21 reliability for summed and scaled scores for polytomously-scored tests. *Applied Measurement in Education, 34*(2), 139-149. https://doi.org/10.1080/08957347.2021.1890740

Al-zayat, F. & Almehrizi, R. S. (2011). *Technical manual for Gulf Multiple Mental Abilities Scale*. Arab Bureau of Education for the Gulf States: Riyadh.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Barbera, J., Naibert, N., Komperda, R., & Pentecost, T. (2021). Clarity on Cronbach's alpha use. *Journal of Chemical Education, 98*, 257-258. https://doi.org/10.1021/acs.jchemed.0c00183.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*, 1-21. https://doi.org/10.1080/08957347.2011.532417

Brennan, R. L. & Lee, W. (1999). Conditional scale-score standard error of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement, 59*, 8-24. https://doi.org/10.1177/0013164499591001

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. https://doi.org/10.1007/BF02310555

El-Keshky, M., & Emam, M. (2015). Emotional and behavioral difficulties in children referred for learning disabilities from two Arab countries: A cross-cultural examination of the Strengths and Difficulties Questionnaire. *Research in developmental disabilities*, *36*, 459-469. https://doi.org/10.1016/j.ridd.2014.10.039

Ellis, J.L. (2021). A test can have multiple reliabilities. *Psychometrika*. https://doi.org/10.1007/s11336-021-09800-2.

Emam, M. M., Almehrizi, R., Omara, E., & Kazem, A. M. (2021). Screening for learning disabilities in Oman: confirmatory factor analysis of the Arabic version of the learning disabilities diagnostic inventory. *International Journal of Developmental Disabilities*, *67*(6), 435-445.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44*(4), 883-891.

Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education, 11*(2), 159-177.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74*, 121-135. https://doi.org/10.1007/s11336-008-9098-4

Hammill, D. D., & Bryant, B. R. (1998). *LDDI: Learning disabilities diagnostic inventory (Examiner Manual)*. Austin: Pro-Ed.

Hammill, D. D., & Bryant, B. R. (2011). *LDDI: Learning disabilities diagnostic inventory (Examiner Manual-2nd Ed.)*. Austin: Pro-Ed.

Kolen, M. J. & Lee, W. (2011). Psychometric properties of raw and scaled scores on mixed-format tests. *Educational Measurement: Issues and Practice, 30*, 15-24. https://doi.org/10.1111/j.1745-3992.2011.00201.x

Kolen, M. J., Hanson, B. A., & Brennan, R. L.(1992). Conditional standard errors of measurement for scaled scores. *Journal of Educational Measurement, 29*, 285-307. https://www.jstor.org/stable/1435086

Kolen, M. J., Wang, T., & Lee, W. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing, 12*, 1-20. https://doi.org/10.1080/15305058.2011.617476

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scaled scores using IRT. *Journal of Educational Measurement, 33*, 129-140. https://doi.org/10.1111/j.1745-3984.1996.tb00485.x

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31*, 255-274. https://doi.org/10.1177/0146621606294206

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*, 375-385. https://doi.org/10.1177/014662169802200407

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). Rejoinder: The Future of Reliability. *Psychometrika*. https://doi.org/10.1007/s11336-021-09807-9