A Comparison of Modern and Popular Approaches to Calculating Reliability for Dichotomously Scored Items

Sébastien Bélard and Carl F. Falk

Abstract
Recent work on reliability coefficients has largely focused on continuous items, including critiques of Cronbach’s alpha. Although two new model-based reliability coefficients have been proposed for dichotomous items (Dimitrov, 2003a,b; Green & Yang, 2009a), these approaches have yet to be compared to each other or other popular estimates of reliability such as omega, alpha, and the greatest lower bound. We seek computational improvements to one of these model-based reliability coefficients and, in addition, conduct initial Monte Carlo simulations to compare coefficients using dichotomous data. Our results suggest that such improvements to the model-based approach are warranted, while model-based approaches were generally superior.

Keywords
reliability, factor analysis, item response theory, classical test theory, dichotomous data

Introduction
Reliability is a fundamental element of scientific measurement (Brennan, 2001; Klaassen & van Peppen, 1986). In psychometric applications, reliability is sometimes described as a way to quantify measurement error: “a unitless index assigned to a measurement instrument, taking its value on the interval [0, 1], for the extent to which the measurement instrument is free from error, with the values 0 and 1 corresponding to the extreme cases of ‘pure error’ and ‘no error,’ respectively” (Li, 2003, p. 91).

There are traditionally three strategies to operationalize reliability; the first two implicate the idea of reliability across time (called test-retest procedure) and across alternative versions of the same test. In this article, we consider the internal consistency strategy, which requires only a single...
administration of the test and of which alpha (α hereafter; Cronbach, 1951) is a classic example. We propose that the internal consistency approach is the most widely used in educational and psychological measurement.

The proper assessment of the reliability of items from a dichotomous test (e.g., correct and incorrect answers) has been neglected: "the majority of the work on it has been concerned with reliability of multiple-item measuring instruments consisting of continuous components" (Raykov et al. 2010, p. 265). Indeed, many authors have recommended against α as a measure of reliability (e.g., cf. Savalei & Reise, 2019; Sijtsma, 2009; McNeish, 2018; Raykov & Marcoulides, 2019), with some suggesting alternatives such as omega coefficients (McDonald, 1985, 1999) and the greatest lower bound (Jackson & Agunwamba, 1977). These articles are largely based on research amongst continuous items.

While many early studies of reliability were rooted in classical test theory (CTT), several researchers have discussed reliability for dichotomous items from a model-based perspective. Bartholomew and colleagues (Bartholomew et al., 1993) defined a reliability coefficient based in part on the two-parameter logistic model (2PLM, Birnbaum, 1968), which may be thought of as a non-linear factor analysis model for dichotomous data (Bartholomew et al., 2008). The 2PLM is one item model under a broader item response theory (IRT) framework, which usually views reliability as conditional upon the level of the latent construct. Bartholomew and colleagues’ work instead allowed computation of a single reliability coefficient as an overall property of test scores. A single reliability coefficient derived from an IRT model allows applied researchers the convenience of reporting a single coefficient, and is in a similar vein as prior work on marginal reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984).

While Bartholomew et al.’s work assumed the researcher created a weighted sum of item scores, it may be more common to compute a test score composite as an unweighted sum. Dimitrov (2003a, 2003b) presented a strategy to define and estimate a single reliability coefficient under these conditions while using the 2PLM. The total score (also referred to as the test score composite, sum score, or number correct score) when computed as an unweighted sum is defined as the following

\[ Y = Y_1 + Y_2 + \cdots + Y_I, \]

where \( Y_1 \) through \( Y_I \) are dichotomous responses (scored 0 and 1) to \( I \) items. This latter case is appropriate when the total of correctly answered items is used as the test score on a multiple choice exam. Although use of this sum score differs from typical IRT or factor analytic scoring methods, this manuscript mainly concerns the reliability of this scoring approach. We argue that this is the most ubiquitous type of data in educational testing, and is common on some psychological tests (e.g., Raskin & Hall, 1979).

The goal of Dimitrov (2003a, 2003b) was to define and estimate reliability with its classical definition, defined as a ratio of true score variance to total variance

\[ \rho_{yy} = \frac{\text{Var}(T)}{\text{Var}(Y)} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E)} = 1 - \frac{\text{Var}(E)}{\text{Var}(Y)} \]  \quad (1)  

where the variance of the total score is decomposed into true score variance, \( \text{Var}(T) \), and error variance, \( \text{Var}(E) \), since \( \text{Var}(Y) = \text{Var}(T) + \text{Var}(E) \) (Lord & Novick, 1968). To further develop this method and make it accessible, Raykov et al. (2010) proposed a strategy to both compute confidence intervals for reliability and estimates of change in reliability when an item is deleted, also using item parameters from the 2PLM. This reliability estimate is based on a function of parameters and standard errors using the Delta method and was implemented with Mplus (Muthén & Muthén, 1998-2017). Given the equivalence of factor analysis of categorical items and some
item response theory models (e.g., Kamata & Bauer, 2008; Takane & de Leeuw, 1987), it may not be surprising that a similar strategy for estimating reliability was developed by Green and Yang (2009b) for nonlinear factor analysis models. This latter approach may very well be equivalent to that of Raykov et al. (2010), but small differences in reliability estimates may occur if the model is estimated in a different way (e.g., Bolt, 2005).

Given these recent developments and the continued use of overall reliability coefficients, including for dichotomous items, more advice is required for applied researchers who may implement such a coefficient. What evidence is there that either approach by Dimitrov (2003a, 2003b) or Green & Yang (2009b) performs better relative to each other, versus Cronbach’s $\alpha$, model-based reliability coefficients from linear structural equation modeling, or any other host of coefficients that could be obtained from the literature (e.g., Guttman’s lambda or the great lower bound (GLB))? Dimitrov (2003a, 2003b) and Raykov et al. (2010) do not present results of simulations, although there are explicit comparisons with $\alpha$ in empirical examples, with $\alpha$ yielding lower estimates of reliability, possibly due to bias and underestimation. Green & Yang (2009b) conducted a small set of simulations in which they suggest that their approach is superior to model-based approaches from linear structural equation modeling, and to $\alpha$. To our knowledge, other recent comparisons with dichotomous items tend to focus exclusively on empirical data (e.g., Deng & Chan, 2017; Napolitano et al., 2013). We are therefore unaware of any study that focuses specifically on dichotomous data that has compared these recently developed model-based approaches to each other, or to any other popular alternatives.

Our main goal is to compare the aforementioned coefficients, with a focus on additionally improving Dimitrov’s (2003a, 2003b) reliability method. We note that Dimitrov’s approach, along with that of Green & Yang (2009b) have not been previously compared under idealized conditions (i.e., a correctly specified model). In the remainder of this article, we present two simulation studies: Study 1 investigates the accuracy of the Dimitrov (2003a, 2003b) analytic approach versus a quadrature-based approach for computing classical reliability from 2PLM item parameters and a known population model. Study 2 then employs simulations to compare the precision of this coefficient with that of Green and Yang, and other popular CTT-based and model-based reliability coefficients. This study is interesting in part because we study the behavior of reliability coefficients with dichotomous data sets as opposed to generating coefficient values from factor analysis model for continuous data (e.g., Trizano-Hermosilla & Alvarado, 2016).

**Study 1**

In studying the accuracy of reliability estimates under the case of dichotomous items, we first need a viable way to define the true, or population-level reliability in a classical sense. Dimitrov (2003a, 2003b; see also May & Nicewander, 1994), provides both the underlying analytical formulas and computational approach for the true reliability of total scores whose item responses are generated by a 2PLM for all $I$ items ($i = 1, 2, \ldots, I$)

$$P_i(\theta) = \frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]},$$

with $D$ equal to the scaling constant that puts the slope, $a_i$, and difficulty, $b_i$, on the approximate same scale as the normal ogive variant of the model (Lord & Novick, 1968). Conceptually, $P_i(\theta)$ provides the probability of obtaining a score of “1” on the item, conditional on the latent trait, $\theta$. To reiterate the analytical expressions for computation of reliability, the error variance for item $i$ is the following
\[ \text{Var}(E_i) = \int_{-\infty}^{\infty} P_i(\theta) [1 - P_i(\theta)] \phi(\theta) d\theta, \]  
(3)

where the integral is across the latent distribution with \( \phi(\cdot) \) its density function. Accumulating over \( I \) items, the entire error variance of the test is the following

\[ \text{Var}(E) = \sum_{i=1}^{I} \text{Var}(E_i), \]  
(4)

The total true-score variance of the test is the following

\[ \text{Var}(T) = \int_{-\infty}^{\infty} \left[ \sum_{i=1}^{I} P_i(\theta) \right]^2 \phi(\theta) d\theta - \left[ \int_{-\infty}^{\infty} \sum_{i=1}^{I} P_i(\theta) \phi(\theta) d\theta \right]^2, \]  
(5)

where \( P_i(\theta) \) is the conditional probability of a correct answer on item \( i \). Finally, Equations (4) and (5) can then be used to define the classic reliability of the test, \( \rho_{yy} \), via equation (1). Implicit in this approach is that the true score variability is due entirely to the underlying latent trait, \( \theta \), and any remaining item-specific variance is considered error.

As opposed to evaluating the integrals in 3 and 5 directly, Dimitrov (2003a, 2003b) proposed an analytical approximation based in part on the error function, we call “\( \text{erf}(\cdot) \)” (see Hastings, 1955). It is claimed that the error function “can be evaluated with an absolute error smaller than 0.0005” (Dimitrov, 2003a; p. 443) to reproduce the marginal probabilities of response to any particular item while assuming a known normal population distribution for the latent trait.

Equations utilizing this approach were then developed to allow the calculation of true reliability. For instance, the marginal probability of a “1” response to item \( i \) is \( P_i(\theta) \) weighted by the density of the latent trait distribution, and then integrating over \( \theta \)

\[ \pi_i = \int_{-\infty}^{\infty} P_i(\theta) \phi(\theta) d\theta, \]  
(6)

Under the 2PLM, this quantity can also be approximated using the error function

\[ \pi_i \approx 1 - \frac{\text{erf}(Z_i)}{2}, \]  
(7)

where \( Z_i = \frac{a_i b_i}{\sqrt{2(1+\sigma_i^2)}} \) and \( \text{erf}(Z_i) = 1 - (1 + m_1 Z_i + m_2 Z_i^2 + m_3 Z_i^3 + m_4 Z_i^4)^{-1} \) assuming \( m_1 = 0.278393, m_2 = 0.230389, m_3 = 0.000972, \) and \( m_4 = 0.078108 \). In addition, when \( Z_i < 0 \), \( \text{erf}(-Z_i) = -\text{erf}(Z_i) \).

Then, the error variance for a particular item can also be approximated by

\[ \text{Var}(E_i) \approx m_i \exp \left[ -0.5 \left( \frac{b_i}{d_i} \right)^2 \right], \]  
(8)

where \( m_i = 0.2646 - 0.118a_i + 0.0187a_i^2 \) and \( d_i = 0.7427 + 0.7081/a_i + 0.0074/a_i^2 \).

Both of these quantities can be used to approximate the true score variance for the number-right score,

\[ \text{Var}(T) \approx \sum_{i=1}^{I} \sum_{j=1}^{J} \sqrt{[\pi_i(1 - \pi_i) - \text{Var}(E_i)] [\pi_j(1 - \pi_j) - \text{Var}(E_j)]}, \]  
(9)

the error variance of the test, and finally the reliability of the test, \( \rho_{yy} \), using (1).
Although this technique may be computationally fast, it has not been evaluated against a more accurate approach. For example, it is not too difficult to evaluate the integral in (3) directly using rectangular quadrature

\[ \text{Var}(E_i) \approx \sum_{q=1}^{Q} P_i(X_q) \left( 1 - P_i(X_q) \right) W_q, \]  

(10)

where \( W_q = \phi(X_q)/\sum_{q=1}^{Q} \phi(X_q) \) are normalized quadrature weights. The integrals in (5) can also be evaluated in this way

\[ \text{Var}(T) \approx \sum_{q=1}^{Q} \left[ \sum_{i=1}^{I} P_i(X_q) \right]^2 W_q - \left[ \sum_{q=1}^{Q} \sum_{i=1}^{I} P_i(X_q) W_q \right]^2, \]  

(11)

If greater precision is desired, the number of weights or their spacing can be adjusted. A strategy used in this Study was to use 101 equally spaced quadrature nodes between -6 and 6.

Evaluation of any discrepancies between what we will call the analytical-based approach of Dimitrov (2003a, 2003b) and this quadrature-based approach is important, since either approach can be used in practice along with item parameters (\( a_i \) and \( b_i \) for all items) to define population reliability, or with estimated parameters as a model-based estimate of reliability. Important to note, both of these approaches make the same assumptions and differ only in the computation approach.

**Method**

We applied a simulation-based approach to study the accuracy of the analytic and quadrature-based approaches for computing classical reliability from 2PLM item parameters. Since true reliability is not known, and must be approximated by either approach, we instead describe the logic of an alternative criterion. Suppose that two variables, \( X \) and \( Y \), are error-prone measures of \( \delta_X \) and \( \delta_Y \), respectively. What we desire is the true correlation between these variables, \( r_{\delta_X \delta_Y} \), but can only observe the correlation between the error-prone variables, \( r_{XY} \). The well-known disattenuation formula is the following (Spearman, 1904)

\[ r_{\delta_X \delta_Y} = \frac{r_{XY}}{\sqrt{\rho_{xx} \rho_{yy}}}, \]  

(12)

which allows us to calculate the true correlation, \( r_{\delta_X \delta_Y} \), by adjusting \( r_{XY} \) for the reliability of sum scores \( X \) and \( Y \), which we denote \( \rho_{xx} \) and \( \rho_{yy} \), respectively.

Now, suppose that we can simulate both error-free, \( \delta_X \) and \( \delta_Y \), and error-prone sum score variables, \( X \) and \( Y \), such that the true correlation \( r_{\delta_X \delta_Y} \) is known, but the reliabilities of the individual variables, \( \rho_{xx} \) and \( \rho_{yy} \), are not known. Based on such data, it should be possible to evaluate the accuracy of different ways of computing \( \rho_{xx} \) and \( \rho_{yy} \). Specifically, along with an estimate of \( r_{XY} \), a method that more accurately computes \( \rho_{xx} \) and \( \rho_{yy} \) should allow better recovery of the true correlation, \( r_{\delta_X \delta_Y} \), when computed using (12).

The general strategy for this study was to simulate latent variables, \( \delta_X \) and \( \delta_Y \), with both unit variance and from a multivariate normal distribution with a known population correlation. These represent error-free versions of two variables. Based on a set of items for each latent variable, and a set of known item parameters for the 2PLM, item responses to \( I \) items per latent construct \( (X_1, \ldots, X_I \text{ and } Y_1, \ldots, Y_I) \) were then generated for each true value of the latent variable. Error-prone versions of \( \delta_X \) and \( \delta_Y \) were then constructed based on the idea of sum score, respectively: \( X = \sum_{i=1}^{I} X_i \) and \( Y = \sum_{i=1}^{I} Y_i \). The final step then involved calculating an estimate of the error-prone correlation between \( X \) and \( Y \), \( \hat{r}_{XY} \), and then adjusting this correlation estimate by the reliabilities,
and $\rho_{xy}$, to arrive at an estimate of $\tilde{r}_{\delta x \delta y}$. Since the true item parameters are known, we directly computed the reliability of the sum scores using either quadrature or analytical approximations.

Study 1 was a 3 (correlation method: $\tilde{r}_{\delta x \delta y}$ with reliabilities based on either quadrature or analytical formulas, and $\tilde{r}_{xy}$) $\times$ 4 (true correlation, $r_{\delta x \delta y} = .3,.5,.7,.9$) $\times$ 13 (number of items per construct: 10 to 70 in increments of five) factorial design. Each combination of true correlation with their specific number of items, was generated; that being 5000 datasets with $N = 500$ respondents each. All three correlation estimates were then computed and recorded for each dataset. Latent trait scores were generated using the mvrnorm function in the MASS package (Venables & Ripley, 2002). True item parameters were drawn randomly across items and datasets, with $a_i \sim \log N(-.75, .75^2)$ and $b_i \sim N(0, .75^2)$. Item responses were generated using standard techniques for the 2PLM (we used the cacIRT package; Lathrop, 2015).

**Results and Discussion**

In this segment, we report the average correlation estimates across all 5000 datasets in each cell of the experimental design. We graphically depict these correlations for the full range of number of items per construct, for each of the three correlation estimation methods, and for all true

![Figure 1](image_url)

**Figure 1.** Estimated correlation where the true correlation is $r_{\delta x \delta y} = .3$ (top-left), $r_{\delta x \delta y} = .5$ (top-right), $r_{\delta x \delta y} = .7$ (bottom-left) and $r_{\delta x \delta y} = .9$ (bottom-right).
correlation values (Figure 1 top-left: $r_{\delta_x\delta_y} = .3$, top-right: $r_{\delta_x\delta_y} = .5$, bottom-left: $r_{\delta_x\delta_y} = .7$, and bottom-right: $r_{\delta_x\delta_y} = .9$).

The true correlation is represented as a horizontal dashed line. The sum score correlation was error-prone and systematically underestimated the true correlation, but improved as the number of items (and therefore reliability) increased. Disattenuated correlations using the analytical approximation for reliability also resulted in an underestimate of the true correlation, but also improved as the number of items increased. For example, in the case where the true correlation was .7 and the number of items was 10 per construct, the average estimated correlation for this method was approximately .65. The quadrature-based approximation was barely affected by the number of items, and resulted in a milder underestimate of the true correlation. This method never resulted in an average disattenuated correlation that differed from the true correlation by more than .006 in absolute value.

Since smaller values of reliability in the disattenuation formula tend to result in larger adjusted correlations, this pattern of results suggests that the analytical approximation would produce a mild overestimate of true reliability. The values of reliability, $\rho_{xx}$ and $\rho_{yy}$, produced by the analytical approximation, are too large to fully disattenuate the error-prone correlation, $r_{xy}$. We therefore conclude that the quadrature-based method for computing reliability is a better benchmark for the study of reliability estimates that we conduct in the following study.

**Study 2**

In order to complement our previous findings, this second study is conducted to compare reliability coefficients with dichotomous items using a simulation study. We focus on reliability coefficients that are popular both historically and currently, as well as include recently developed model-based coefficients. We mainly divide such reliability coefficients into those based on CTT (or observed scores) and those based on a measurement model.

**CTT-Based Coefficients**

From Guttman’s (1945) six lower bounds of reliability, $\lambda_3$ is the same as the popular Cronbach’s $\alpha$ (1951), which is also a generalization of a coefficient found in Kuder & Richardson (1937). Mathematically,

$$\lambda_3 = \alpha = \left( \frac{I}{I-1} \right) \left[ 1 - \frac{\sum_{i=1}^{I} \text{Var}(Y_i)}{\text{Var}(Y)} \right],$$

(13)

where $\text{Var}(Y_i)$ is the variance of item $i$ and $\text{Var}(Y)$ is the variance of the sum score composite, $Y$. Under a congeneric model, $\alpha$ is expected to underestimate reliability (e.g., Raykov, 1997), but may also overestimate if errors are correlated (Raykov, 2001). Interestingly, the usefulness of this coefficient was recently discussed with some authors suggesting to abandon $\alpha$ (e.g., Dunn et al., 2013; McNeish, 2018) others argued that this coefficient is still useful (e.g., Raykov & Marcoulides, 2019).

Sijtsma (2009) suggested use of the great lower bound (GLB) as an alternative to coefficient $\alpha$. For the GLB, we first present a matrix decomposition of the $I$ observed scores for the test (e.g., Jackson & Agunwamba, 1977)

$$\Sigma_Y = \Sigma_T + \Sigma_E,$$

(14)
where the $I \times I$ covariance matrix, $\Sigma_Y$, is decomposed into true score covariances, $\Sigma_T$, and error covariances, $\Sigma_E$, noting that the latter is usually assumed diagonal and contains only error variances. Since the right-hand decomposition of (14) is unknown, the elements of this equation must be estimated. An alternative decomposition is provided by Bentler & Woodward (1980; see also Bentler, 2009) and may be more useful for understanding why the GLB is considered a “lower bound” to reliability. The true score covariance matrix is further decomposed, $\Sigma_T = \Sigma_C + \Sigma_S$, such that

$$\Sigma_Y = \Sigma_C + \Sigma_S + \Sigma_E,$$  \hspace{1cm} (15)

Note that $\Sigma_C$ is now the covariance matrix among variables due to some common source of variability and $\Sigma_S$ is a matrix that is usually diagonal and contains item-specific variances. In practice, it is easier to think of $\Sigma_C$ as covariance among variables that is due to some underlying cause or construct that the test creator is trying to measure.

Finally, the GLB is computed as follows

$$GLB = 1 - \frac{\text{tr}(\Sigma_E)}{\text{Var}(Y)},$$  \hspace{1cm} (16)

where $\text{tr}(\Sigma_E)$ is the trace of the inter-item error covariance matrix. In comparing (16) to equation (1), we see that $\text{tr}(\Sigma_E) = \text{Var}(E)$, provided that the matrix is diagonal. Important for understanding this coefficient is that $\Sigma_C$ may be considered to contain reliable, though item-specific variance. However, it may be impossible to empirically distinguish between $\Sigma_S$ and $\Sigma_E$ in a single test administration. Regardless, any estimate of $\text{tr}(\Sigma_E)$ may actually be an estimate of $\text{tr}(\Sigma_S + \Sigma_E)$, and reliability is arguably underestimated. This occurs since the GLB in (16) becomes smaller when the error variances in $\Sigma_E$ are larger. The GLB is therefore described as the “worse-case scenario” (Sijtsma, 2009; Sijtsma & van der Ark, 2017) as actual computation of (16) is done by maximizing $\text{tr}(\Sigma_E)$ subject to the constraints that $\Sigma_E$ is nonnegative and $\Sigma_T$ is positive semidefinite.

**Model-Based Reliability**

Use of a measurement model often entails additional assumptions regarding functional form or the distribution of the variables, but can sometimes offer greater flexibility or additional features (e.g., multidimensional models, adaptive test assembly) if such assumptions are met. In what follows, we will shortly discuss approaches based on both linear and non-linear models.

**Linear Model-Based**

**Omega** ($\omega$), is perhaps the most commonly reported linear model-based coefficients and was developed by McDonald (1985, 1999), yet has its roots in work by Joreskog (1971) and is sometimes called coefficient rho, $\rho$ (e.g., Raykov & Shrout, 2002). We use the term “linear” here to indicate that estimates of this reliability coefficient can be derived from parameter estimates of a linear factor analysis model. The common factor model with a single underlying factor takes the following form for item $i$

$$Y_i = \mu_i + \lambda_i \theta + \epsilon_i,$$  \hspace{1cm} (17)

where $\mu_i$ is an intercept, $\lambda_i$ is a slope known as a factor loading, $\theta$ is a latent trait and $\epsilon_i$ is the random residual of the model (with $\text{Var}(\epsilon_i) = \psi_i^2$). McDonald (1999) equates the variability due to $\theta$ as a common source of reliable variance in line with the decomposition in (16). In the case of a
one-factor model and assuming uncorrelated errors, true score (or common) variance is approximated via the sum of squared loadings, $\text{Var}(T) \approx (\sum_{i=1}^{I} \lambda_i)^2$, error variance is approximated by $\text{Var}(E) = \sum_{i=1}^{I} \psi_i^2$, and results in an estimate of reliability

$$\omega = \frac{(\sum_{i=1}^{I} \lambda_i)^2}{(\sum_{i=1}^{I} \lambda_i)^2 + \sum_{i=1}^{I} \psi_i^2}$$

(18)

which can be related to the second equality of equation (1), even if there are conceptual differences between both approaches. According to Maydeu-Olivares et al. (2007): “If a good-fitting model can be found, the use of a model based reliability estimate is clearly the best option. For instance, if a one-factor model is found to fit the data well, then the reliability of the test score is given by coefficient $\omega$, and the applied researcher should employ this coefficient” (p. 172). However, we have to mention there are many ways to estimate $\omega$ (Revelle & Condon, 2019). In this article, we will focus only on $\omega_{Total}$.

There are also several ways to estimate $\lambda_i$ and $\psi_i^2$. By default, the psych package (Revelle, 2019) uses “minres” (ordinary least squares), which we study ($\omega$) in addition to estimation using confirmatory factor analysis estimated via normal theory maximum likelihood (CF$A$). However, this is somewhat theoretically questionable with dichotomously scored data as it may be more reasonable to assume that there is a nonlinear relationship between the underlying common factor and the item responses, which we discuss in the next section.

**Non-Linear Model-Based**

We are aware of two general approaches whereby an estimate of reliability is obtained by an IRT or nonlinear factor analysis model. We have already discussed work by Dimitrov (2003a, 2003b) and how true reliability for a test consisting of dichotomous items can be defined in terms of item parameters for the 2PLM. Raykov et al. (2010) took this one step further and reasoned that consistent estimates of item parameters should yield a consistent estimate of reliability. Therefore, it is feasible to estimate all item parameters using the 2PLM and then utilize either analytical (which we call DA) or quadrature-based equations (DQ) to estimate reliability using Equations (7–11). In our simulations, we take both of these approaches after estimating the 2PLM with the mirt package (Chalmers, 2012) in R.

In addition, Green and Yang (2009b) developed a method based on a nonlinear structural equation model (GY). These authors derived an estimate of reliability by defining reliability as the correlation between a test and its parallel form. The resulting equations are in terms pertaining to the item parameter estimates from a nonlinear factor analysis model. In supposing that the latent trait is linearly related to some underlying item response, and that this response variable is dichotomized, the model takes the underlying variable approach. To estimate the model, a limited information estimation approach is employed whereby item thresholds are estimated, followed by polychoric correlations, and then item parameters (see Bolt, 2005). The resulting thresholds and factor loadings can be used to estimate the reliability between a sum score composite of the observed variables and their parallel form, which, in theory, should be the same as classical definitions of reliability. It is well-known that such a factor analysis model of dichotomous items is equivalent to an IRT model such as the 2PLM, except a probit link function is often used in place of a logistic function (Kamata & Bauer, 2008; Takane & De Leeuw, 1987). We therefore speculate that this approach is likely equivalent (or nearly so) to the approach presented by Raykov et al. (2010). Since the equations and derivations for this approach are quite long and require another
model parameterization, we refer the reader to the original article for more details (e.g., Green & Yang, 2009b; Equation (21), p. 160).

**Some Theoretical Expectations**

By using data generated under a 2PLM, we may get a sense of the behavior of such reliability estimates in relation to realistic slopes and discrimination parameters. We also presume that the most realistic scenario is that item parameters such as the discrimination, \( a_i \), and difficulty, \( b_i \), vary across items. This has several implications for the expected performance of the just presented reliability estimates.

First, the relationship between the latent trait and the underlying variable is not the same for each variable, and the variance of each variable may be different. In such a case, the so-called congeneric model results (Raykov, 1997). Under this model, coefficient \( \alpha \) is expected to be an underestimate of reliability. Although the GLB is theoretically a better estimate, some bias and instability has been found in previous research (Sijsma & van der Ark, 2017).

Next, many authors promote \( \omega \) because this coefficient was found to be superior to \( \alpha \) (e.g., Dunn et al., 2013; Maydeu-Olivares & Hartmann, 2007; McDonald, 1985; Trizano-Hermosilla & Alvarado, 2016). Even though linear model-based estimates of reliability such as \( \omega \) are derived from what may be considered an incorrect model for the data, to the extent that a linear function of the latent trait still provides a reasonable estimate of the variability due to the trait, \( \omega \) may be a decent approximation of reliability.

Finally, nonlinear model-based estimates of reliability are theoretically the most correct. However, in finite samples and tests, there may be small differences in both bias and efficiency of item parameter estimates used in IRT and factor analysis of categorical data (e.g., Forero & Maydeu-Olivares, 2009; Rhemtulla et al., 2012), which may lead to small differences in coefficient performance. As test length and sample size increases, however, we would expect that such approaches would perform the best.

In the following simulation study, we compare these coefficients under a variety of sample sizes and test lengths, and vary the true item parameters generated across replications.

**Method**

The design crossed 3 test lengths (\( I = 15, 40, 65 \)) and 5 sample sizes (\( N = 100, 300, 500, 1000, 3000 \)). For each combination, 1000 datasets were generated for a total of \( (3\times5)\times1000 = 15000 \) datasets. For each dataset, item parameters were drawn randomly from the following distributions: \( a_i \sim \text{log}N(-.75,.25^2) \), \( b_i \sim N(0,.75^2) \). The latent trait was also normally distributed, \( \theta \sim N(0,1) \), and such item parameters and latent trait scores were used to generate responses to the dichotomous items using the sim function from the caicIRT R package (Lathrope, 2015).

The coefficients under consideration are presented in the next table (Table 1), and have already been discussed in the previous section, including also \( DA \) using true item parameters (\( DA_{\text{True}} \)). Example code for computing all coefficients appears in Supplementary Materials.

We compared true reliability to the value of every coefficient by reporting the root mean square error

\[
RMSE = \sqrt{\frac{\sum_{s=1}^{1000} (\hat{\rho}_s - \rho_{\text{True}})^2}{1000}}, \tag{19}
\]

and bias

\[
\text{Bias} = \frac{\sum_{s=1}^{1000} (\hat{\rho}_s - \rho_{\text{True}})}{1000}.
\]
\[
\text{Bias} = \sum_{s=1}^{1000} \frac{\left(\hat{\rho}_s - \rho_{\text{True}}\right)}{1000},
\]

where \(\hat{\rho}_s\) is the value of an estimated coefficient (e.g., \(DQ\), \(\alpha\) or \(\omega\)) for simulated dataset \(s\) and the true reliability \(\rho_{\text{True}}\) is computed using (10) and (11).

**Results**

A similar pattern of results emerged across both sample size and the number of items. We describe these results as concisely as possible, while mentioning general patterns across these manipulated factors.

**RMSE**

Not surprisingly, Table 2 shows that RMSE became smaller when the number of items and the sample size increased. The top-three reliability coefficients with the smallest RMSE are \(DQ\), \(CFA\) and \(\omega\), respectively, for 15 items; \(\omega\), \(DQ\) and \(CFA\), respectively, for 40 items; and \(\omega\), \(DQ\) and \(CFA\), respectively, for 65 items. \(\alpha\) and \(GY\) were close behind in both cases. Differences among these approaches appeared negligible, and it was interesting that these coefficients were not very sensitive to sample size.

At the opposite end of performance, \(GLB\) presented the largest RMSE and appeared to be strongly affected by sample size. Note how the data points for sample size vary greatly for the \(GLB\) in any given case. For example, in the case of a 15 item test, \(GLB\) presented an RMSE of 5.388 for \(N = 100\) and .94 when \(N = 3000\).

| Table 1. Coefficients Under Investigation in Study 2 |
|------------------------------------------|
| Coefficient name’s | Information | Source |
| DATrue | Method developed by Dimitrov (2003a) based on analytic approximation using the true \(a_i\) and \(b_i\) parameters | Custom code in supplementary materials |
| DQ | Method developed by Dimitrov (2003a) based on rectangular quadrature used after estimating item parameters with the 2PLM | Custom code in supplementary materials |
| DA | Method developed by Dimitrov (2003a) based on an analytical approximation used after estimating item parameters with the 2PLM | Custom code in supplementary materials |
| \(\alpha\) | Cronbach’s \(\alpha\) or Guttman’s \(\lambda_3\) | alpha function from psych R package | Revelle (2019) |
| GLB | \(\omega_{\text{Total}}\) based on 1 dimension using pearson correlation | glb.algebraic function from psych R package | omega function from psych R package |
| \(\omega\) | GLB based on an algebraic procedure | | |
| CFA | Model-based reliability from loadings and errors variances of a single-factor CFA | Custom code using lavaan function from lavaan R package (Rosseel, 2012) |
| GY | Green and yang coefficient (2009b) based on WLS estimator | Custom code using lavaan function from lavaan R package and sirt R package (Robitzsch, 2019) |
In line with the previous results, bias (Table 3) became smaller when the number of items and sample size increased. In addition, our results revealed that $DATrue$, $DA$ and $GLB$ systematically presented a positive bias. That said, the bias was sometimes high, and primarily under conditions with few items. On the other hand, $\alpha$ and $CFA$ all presented a systematic negative and small bias over the simulations.

The coefficients with the smallest level of bias were where all model-based coefficients $DQ$, $\omega$ and $CFA$. Again, the magnitude of bias did not seem to be affected by the sample size or the number of items. At the opposite end, $GLB$ coefficients were those with the largest bias, which is particularly true when the sample size decreased and may explain part of the reason $GLB$ coefficients had unstable RMSE.

### Table 2. RMSE.

| N  | DATrue | DQ | DA | $\alpha$ | GLB | $\omega$ | CFA | GY |
|----|--------|----|----|---------|-----|--------|-----|----|
| 100| 1.338  | 0.071 | 1.133 | 0.179 | 5.388 | 0.088 | 0.020 | 0.285 |
| 300| 1.169  | 0.019 | 1.071 | 0.155 | 2.803 | 0.035 | 0.059 | 0.121 |
| 500| 1.162  | 0.044 | 1.086 | 0.156 | 2.163 | 0.057 | 0.075 | 0.202 |
| 1000| 1.171 | 0.009 | 1.159 | 0.088 | 1.582 | 0.004 | 0.022 | 0.200 |
| 3000| 1.174 | 0.019 | 1.184 | 0.064 | 0.94 | 0.010 | 0.005 | 0.216 |
| 100 | 0.678 | 0.023 | 0.587 | 0.111 | 3.720 | 0.022 | 0.043 | 0.118 |
| 300 | 0.679 | 0.007 | 0.654 | 0.057 | 2.424 | 0.007 | 0.016 | 0.055 |
| 500 | 0.682 | 0.013 | 0.668 | 0.046 | 1.941 | 0.001 | 0.011 | 0.091 |
| 1000| 0.682 | 0.001 | 0.673 | 0.046 | 1.141 | 0.006 | 0.016 | 0.123 |
| 3000| 0.681 | 0.006 | 0.673 | 0.044 | 0.837 | 0.007 | 0.018 | 0.143 |
| 100 | 0.472 | 0.007 | 0.428 | 0.084 | 2.922 | 0.028 | 0.041 | 0.074 |
| 300 | 0.469 | 0.013 | 0.437 | 0.042 | 2.010 | 0.008 | 0.015 | 0.042 |
| 500 | 0.470 | 0.005 | 0.448 | 0.032 | 1.638 | 0.003 | 0.01 | 0.064 |
| 1000| 0.470 | 0.022 | 0.449 | 0.034 | 1.209 | 0.007 | 0.014 | 0.086 |
| 3000| 0.472 | 0.017 | 0.457 | 0.025 | 0.732 | 0.001 | 0.007 | 0.096 |

### Bias

In line with the previous results, bias (Table 3) became smaller when the number of items and sample size increased. In addition, our results revealed that $DATrue$, $DA$ and $GLB$ systematically presented a positive bias. That said, the bias was sometimes high, and primarily under conditions with few items. On the other hand, $\alpha$ and $CFA$ all presented a systematic negative and small bias over the simulations.

The coefficients with the smallest level of bias were where all model-based coefficients $DQ$, $\omega$ and $CFA$. Again, the magnitude of bias did not seem to be affected by the sample size or the number of items. At the opposite end, $GLB$ coefficients were those with the largest bias, which is particularly true when the sample size decreased and may explain part of the reason $GLB$ coefficients had unstable RMSE.

### Discussion

Several coefficients performed well in our simulation studies. Model-based coefficients $DQ$, $\omega$ and $CFA$ presented the smallest RMSE and bias. Any of these would be a good choice if the user wishes to choose a model-based reliability coefficient, though these may require (testable) model assumptions. Generally speaking, although presenting a single coefficient based on an IRT model can be criticized, $DQ$ may provide a reasonable estimate and future research may show it is a good alternative to so-called Person separation indices (e.g., see Mair et al., 2018; Rasch, 1960).

The $GLB$ was strongly affected by the sample size and is not recommended for a unidimensional test. This is in accordance with Shapiro & Ten Berge (2000). In addition, we found a positive bias as also observed by Ten Berge & Sočan (2004) for this coefficient.

Recent critiques against Cronbach’s $\alpha$ restored the interest of a larger audience for reliability (Dunn, Baguley & Brunsden, 2013; Green & Yang, 2009a; McNeish, 2018; Revelle & Zinbarg, 2009; Sijtsma, 2009; Trizano-Hermosilla & Alvarado, 2016). If a model-based coefficient such as
Table 3. Bias.

| N     | DATrue | DQ   | DA   | α    | GLB   | ω    | CFA   | GY    |
|-------|--------|------|------|------|-------|------|-------|-------|
| 15 items |
| 100   | 0.0481 | 0.0025 | 0.0408 | −0.0065 | 0.1940 | 0.0032 | −0.0007 | 0.0103 |
| 300   | 0.0369 | −0.0006 | 0.0339 | −0.0049 | 0.0886 | −0.0011 | −0.0019 | −0.0038 |
| 500   | 0.0367 | −0.0014 | 0.0343 | −0.0049 | 0.0684 | −0.0018 | −0.0024 | −0.0064 |
| 1000  | 0.0370 | 0.0003 | 0.0367 | −0.0028 | 0.0500 | −0.0001 | −0.0007 | −0.0063 |
| 3000  | 0.0371 | 0.0006 | 0.0375 | −0.0020 | 0.0297 | 0.0003 | −0.0002 | −0.0068 |
| 40 items |
| 100   | 0.0214 | 0.0007 | 0.0186 | −0.0035 | 0.11762 | −0.0007 | −0.0014 | 0.0037 |
| 300   | 0.0215 | 0.0002 | 0.0207 | −0.0018 | 0.0766 | −0.0002 | −0.0005 | −0.0017 |
| 500   | 0.0216 | 0.0004 | 0.0211 | −0.0014 | 0.0614 | −0.0000 | −0.0003 | −0.0029 |
| 1000  | 0.0216 | 0.0000 | 0.0213 | −0.0015 | 0.0447 | −0.0002 | −0.0005 | −0.0039 |
| 3000  | 0.0215 | −0.0002 | 0.0213 | −0.0014 | 0.0265 | −0.0002 | −0.0006 | −0.0045 |
| 65 items |
| 100   | 0.0149 | 0.0002 | 0.0136 | −0.0026 | 0.0924 | −0.0009 | −0.0013 | 0.0024 |
| 300   | 0.0148 | −0.0003 | 0.0138 | −0.0013 | 0.0636 | −0.0003 | −0.0005 | −0.0013 |
| 500   | 0.0149 | −0.0002 | 0.0142 | −0.0010 | 0.0518 | −0.0001 | −0.0003 | −0.0020 |
| 1000  | 0.0149 | −0.0007 | 0.0142 | −0.0011 | 0.0382 | −0.0002 | −0.0005 | −0.0027 |
| 3000  | 0.0149 | −0.0005 | 0.0145 | −0.0008 | 0.0232 | −0.0000 | −0.0002 | −0.0030 |

$DQ$, $ω$ and $CFA$ can be computed, this is preferred. $α$ may be a reasonable alternative only if there are software limitations or if estimation of a model proves difficult (e.g., in situations where the number of items is large or participants are few). Model-based coefficients are more precise and do not require tau-equivalent measurement model assumptions.

Conclusion

This article integrates 2 studies. First, we found that Dimitrov’s (2003a, 2003b) strategy based on quadrature is a better benchmark for the study of reliability. Second, we used a simulation study to analyze the precision of some CTT-based and model-based reliability coefficients. Our results show that Dimitrov’s (2003a, 2003b) quadrature-based strategy, $ω$ based on a single dimension, and $CFA$-based on a single dimension are all the most precise coefficients. At the opposite, $GLB$ is the least interesting coefficient among those under investigation.

This article is obviously not without limitations. As many coefficients studied here have not been previously compared, it was logical to first study them under ideal conditions (i.e., a correctly specified model). It is important in future work to study them under assumption violations (multiple dimensions and locally dependent items; see also van der Ark, van der Palm, & Sijtsma, 2011) and more complex models to understand how well these coefficients behave and where such reliability coefficients may result in larger discrepancies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs
Sébastien Béland https://orcid.org/0000-0002-0324-5076
Carl F. Falk https://orcid.org/0000-0002-4788-7206

Notes
1. We are however aware that many authors criticize the use of sum scores (e.g. Kuhfeld & Soland, 2020; McNeish & Wolf, 2020)
2. On p. 88 of McDonald (1999), for example, the notation “C” is used in place of “T” to indicate common factor, and “U” instead of “E” to indicate a unique part of the item in (1).
3. Several other reliability coefficients were studied for exploratory purposes, but did not perform well or were not as theoretically justified and are omitted due to space constraints: Other variants of Guttman’s coefficient, GLB estimated using factor analytic approach, defaults for $\omega_{Total}$ including three extracted factors or use of polychoric correlations. Results are available upon request.

References
Ark van der, L. A., Daniël, W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. Applied Psychological Measurement, 35(5), 380–392. https://orcid.org/10.1177/0146621610392911
Bartholomew, D. J., Bassin, E. L., & Schuessler, K. F. (1993). Properties of a latent trait reliability coefficient. Sociological Methods and Research, 22(2), 163–192. https://doi.org/10.1111/j.1467-956X.1993.tb00347.x
Bartholomew, D. J., Steele, J., Galbraith, J., & Moustaki, I. (2008). Analysis of multivariate social science data (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b15114
Bentler, P. M. (2009). Alpha, Dimension-Free, and Model-Based Internal Consistency Reliability. Psychometrika, 74(1), 137–143. https://doi.org/10.1007/s11336-008-9100-1
Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability Addison–Wesley. In F. M. M. R. DansLord etNovick (Ed.), Statistical theories of mental test scores. Massachutts.
Bolt, D. M. (2005). Limited- and full-information estimation of item response theory modelsLawrence Erlbaum associates publishers. In A. Maydeu-Olivares, & J. J. McArdle (Eds.), Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald (pp. 27–71).
Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. Journal of Educational Measurement, 38(4), 295–317. https://doi.org/10.1111/j.1745-3984.2001.tb01129.x
Chalmers, P. R. (2012). Mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/bf02310555
Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. Educational and Psychological Measurement, 77(2), 185–203. https://doi.org/10.1177/0013164416658325
Dimitrov, D. M. (2003a). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. Applied Psychological Measurement, 27(6), 440–458. https://doi.org/10.1177/0146621603258786
Dimitrov, D. M. (2003b). Reliability and true-score measures as a function of their Rasch difficulty parameter. *Journal of Applied Measurement*, 4(3), 222–233.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399–412. https://doi.org/10.1111/bjop.12046

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, 14(3), 275–299. https://doi.org/10.1037/a0015825

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, 74(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282. https://doi.org/10.1007/BF0228892

Hastings, C. Jr (1955). *Approximations for digital computers*. Princeton University Press.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic Lower Bounds. *Psychometrika*, 42(4), 567–578. https://doi.org/10.1007/BF02295979

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36(4), 409–426. https://doi.org/10.1007/BF02291366

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, 15(1), 136–153. https://doi.org/10.1080/10705510701758406

Klaassen, K. B., & van Peppen, J. (1986). Reliability in instrumentation. *Journal of Physics E: Scientific Instruments*, 19(8), 577–587. https://doi.org/10.1088/0022-3735/19/8/001

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151–160. https://doi.org/10.1007/bf02288391

Kuhfeld, M., & Soland, J. (2020). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*. https://doi.org/10.1037/met0000367 Advance online publication.

Lathrop, Q. N. (2015). *Classification accuracy and consistency under item response theory*. R package version 1.4. https://CRAN.R-project.org/package=cacIRT.cacIRTURL

Li, H. (2003). The resolution of some paradoxes related to reliability and validity. *Journal of Educational and Behavioral Statistics*, 28(2), 89–95. https://doi.org/10.3102/10769986028002089

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Mair, P., Hatzinger, R., & Maier, M. J. (2018). *eRm: Extended rasch modeling*. http://erm.r-forge.r-project.org/.0.161

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, 12(2), 157–176. https://doi.org/10.1037/1082-989X.12.2.157

May, K., & Nicewander, W. A. (1994). Reliability and information functions for percentile ranks. *Psychometrika*, 31(4), 313–325. https://doi.org/10.1111/j.1745-3984.1994.tb00449.x

McDonald, R. P. (1985). *Factor analysis and related methods*. Lawrence Erlbaum.

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

McNeish, D. (2018). Thanks coefficient alpha, we’ll take it from here. *Psychological Methods*, 23(3), 412–433. https://doi.org/10.1037/met0000144
McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behav Res*, 52(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user’s guide* (8th ed.). Muthén & Muthén.

Napolitano, C. M., Callina, K. S., & Mueller, M. K. (2013). Comparing alternate approaches to calculating reliability for dichotomous data: The sample case of adolescent selection, optimization, and compensation. *Applied Developmental Science*, 17(3), 148–151. https://doi.org/10.1080/10888691.2013.804372

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Danish Institute for Educational Research.

Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports*, 45(2), 590. https://doi.org/10.2466/pr0.1979.45.2.590

Raykov, T. (1997). Scale reliability, Cronbach’s coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32(4), 329–353. https://doi.org/10.1207/s15327906mbbr3204_2

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, 25(1), 69–76. https://doi.org/10.1177/01466216003659417

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, 9(2), 195–212. https://doi.org/10.1207/S15328007SEM0902_3

Revelle, W. (2019). *Psych: Procedures for personality and psychological research*. Northwestern University. https://CRAN.R-project.org/package=psych Version = 1.9.12.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, 31(12), 1395–1411. https://doi.org/10.1037/pas0000754

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rhemtulla, M., Brousseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. https://doi.org/10.1037/a0029315

Robitzsch, A. (2019). *sirt: Supplementary item response theory model*. https://CRAN.R-project.org/package=sirt R package version 3.6-21.

Rosseel, Y (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Savalei, V., & Reise, S. P. (2019). Don’t forget the model in your model-based reliability coefficients: A reply to McNeish (2018). *Collabra: Psychology*, 5(1), 36. https://doi.org/10.1525/collabra.247

Shapiro, A., & ten Berge, J. M. F. (2000). The asymptotic bias of minimum trace factor analysis with applications to the greatest lower bound to reliability. *Psychometrika*, 65(3), 413–425. https://doi.org/10.1007/bf02296154

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., & van der Ark, A. (2017). Reliability. In V. Zeiger-Hill, & T.K. Shackelford (Eds.), *Encyclopedia of personality and individual differences*. Springer.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1), 72–101. https://doi.org/10.2307/1412159
Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408. https://doi.org/10.1007/bf02294363

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613–625. https://doi.org/10.1007/bf02289858

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

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.