Practical Assessment, Research, and Evaluation

Volume 28 Article 10

August 2023

Seeking the Real Reliability: Why the Traditional Estimators of Reliability Usually Fail in Achievement Testing and Why the Deflation-Corrected Coefficients Could Be Better Options

Jari Metsämuuronen
University of Turku, Turku Research Centre for Learning Analytics, and Finnish National Education Evaluation Centre (FINEEC)

Seeking the Real Reliability: Why the Traditional Estimators of Reliability Usually Fail in Achievement Testing and Why the Deflation-Corrected Coefficients Could Be Better Options

Cover Page Footnote

Author likes to give sincere thanks for an anonymous reader of constructive suggestions how to amend the text.

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 28 Number 10, August 2023 ISSN 1531-7714

Seeking the Real Reliability: Why the Traditional Estimators of Reliability Usually Fail in Achievement Testing and Why the Deflation-Corrected Coefficients Could Be Better Options

Jari Metsämuuronen, *University of Turku, Turku Research Centre for Learning Analytics*, and *Finnish National Education Evaluation Centre (FINEEC)*

Traditional estimators of reliability such as coefficients alpha, theta, omega, and rho (maximal reliability) are prone to give radical underestimates of reliability for the tests common when testing educational achievement. These tests are often structured by widely deviating item difficulties. This is a typical pattern where the traditional Pearson correlation between items and score (*Rit*) may be radically deflated. Because *Rit* is embedded in the traditional estimators of reliability, this causes deflation in the estimates of reliability, and the magnitude of deflation may be remarkable. Within achievement testing, deflation-corrected estimators of reliability (DCER) would be better options. Instead of *Rit*, DCERs use other estimators of correlation as the linking factor between the item and the score variable that are less prone to deflation. Selecting wisely the linking coefficient, DCERs may give significant advance in estimating the true reliability and true standard error related to the test score.

Keywords: reliability, deflation-corrected reliability, deflation in reliability, deflation in correlation, item-score correlation, coefficient alpha, coefficient theta, coefficient omega, maximal reliability

Introduction

Psychometric testing—including achievement testing in educational settings—has a long history that can be traced to Wilhelm Wundt (1862) in German, Francis Galton (1869) in Great Britain, and J. McKeen Cattell (1886, 1893) in USA (see the history in Gregory, 2004). From early on, three concepts have been of specific interest within test theory—the same are the interest in this article: coefficients of correlation, coefficients of reliability, and the standard error of the measurement (*S.E.m*).

The concept of correlation and, specifically, the product-moment correlation coefficient (PMC) was studied by Karl Pearson (1896 onwards) based on Auguste Bravais' (1844) and Galton's (1889) earlier innovations. Later, Pearson (1903) and Spearman (1904) were the first to point the inaccuracy in PMC and offered the first solutions to the problem of attenuation (see the history in Sackett & Yang, 2000, see also Sackett et. al., 2007; Schmidt & Hunter, 2015). Recently, Metsämuuronen (e.g., 2021a, 2022f) have pointed out that the estimates of correlation by PMC embed not only *attenuation* caused by errors in the measurement modelling but also radical *deflation* due to artificial systemic errors during the estimation (of the concepts, see, Gadermann, Guhn, & Zumbo, 2012; Metsämuuronen, 2022f; Revelle & Condon, 2018; Silver, 2008). When the scales of two variables differ radically from each other, as is always the case with a test item and the score variable, PMC cannot reach the latent correlation but, instead, the deflation approximates 100% when the variance in either variable approximates zero (see examples in

Metsämuuronen, 2022f). In testing settings, radical deflation happens always with items of extreme difficulty level; the estimates of item–total correlation of very easy or very difficult test items are always radically deflated. This issue is discussed later.

The concept of reliability of the test score was first discussed, and the first formulae were derived, by William Brown (before 1910 in his thesis and later in 1910; see Cho & Chun, 2018) and Charles Spearman (1910); the aim was to correct the attenuation in PMC caused by "faulty data". In special cases related to strictly parallel tests, this coefficient of reliability, Brown–Spearman prophecy formula (see the rationale for the non-traditional order of the developers in Cho & Chun, 2018), is still in use, and it is the ancestor of the most widely used estimator of reliability, coefficient alpha (chronologically, Kuder & Richardson, 1937; Jackson & Ferguson, 1941; Cronbach, 1951). Other options for coefficients of reliability within the classical test theory, generalizability theory, and IRT modelling are collected and discussed recently by Metsämuuronen (2022d). Later, reliability turned to be the main concept and tool to quantify the amount of random measurement error that exists in a score variable generated by a compilation of multiple test items as well as to assess the (overall) quality of the measurement (e.g., Gulliksen, 1950). An estimate of reliability serves also in correcting the attenuation in correlations in validity studies and meta-analyses (e.g., Schmidt & Hunter, 2015) and in the estimates of regression or path models (e.g., Cole & Preacher, 2014).

Recently, Metsämuuronen (e.g., 2022a, 2022c, 2022d, 2022e) has discussed a hidden challenge in the estimators of the reliability: the estimates by such traditional estimators as coefficients alpha, theta, omega, and rho (maximal reliability) may be radically deflated because they embed PMC in the form of item–score correlation (*Rit* $=\rho_{iX}$). *Rit* is explicit in the coefficient alpha (see later Eq. (2))—as well as in such preceding estimators as Brown–Spearman prophecy formula (Brown, 1910, Spearman, 1910), Flanagan–Rulon prophecy formula (Rulon, 1939), Kuder–Richardson formulae KR20 and KR21 (Kuder & Richardson, 1937), and Guttman's lambda family (Guttman, 1945)—because the variance of the test score (σ_X^2) is visible in the formula. σ_X^2 on its behalf is inherited from the basic definition of reliability

$$REL = \sigma_T^2/\sigma_X^2 = 1 - \sigma_E^2/\sigma_X^2 \quad (1)$$

e.g., Gulliksen, 1950) where σ_T^2, σ_X^2, and σ_E^2 refer to the variances of the observed score (X) and unobserved true score (T) and error element (E) familiar from their profound relation in testing theory, $X = T + E$. It is known that σ_X^2 can be expressed by item variances σ_i^2 and ρ_{iX}:

$$\sigma_X^2 = \left(\sum_{i=1}^{k} \sigma_i \times \rho_{iX}\right)^2$$

(Lord, Novick, & Birnbaum 1968), where k refers to the number items in the compilation. Then, the coefficient alpha (ρ_α), as an example, can be expressed as

$$\rho_\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right) = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times \rho_{iX}\right)^2}\right) \quad (2)$$

(Lord et al., 1968) where PMC is explicit. Also, in the form of principal component and factor loading (λ_i), *Rit* is embedded in such advanced estimators based on "optimal linear combination" (see Li, 1997) as coefficient theta (ρ_{TH}; Armor, 1974; see also Kaiser & Caffrey, 1965, based on Lord, 1958) based on principal component loadings

$$\rho_{TH} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k}\lambda_i^2}\right), \quad (3)$$

coefficient omega (ρ_ω; Heise & Bohrnstedt, 1970; McDonald, 1970) known also as McDonald's omega total (McDonald, 1999):

$$\rho_\omega = \frac{\left(\sum_{i=1}^{k}\lambda_i\right)^2}{\left(\sum_{i=1}^{k}\lambda_i\right)^2 + \sum_{i=1}^{k}\left(1-\lambda_i^2\right)}, \quad (4)$$

and coefficient rho or maximal reliability (ρ_{MAX}; e.g., Raykov, 1997, 2004 onwards) known also as composite reliability (e.g., Raykov, 1997), Raykov's Rho (e.g., Cleff, 2019), and Hancock's H (Hancock & Mueller, 2001):

$$\rho_{MAX} = \frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(\lambda_i^2/\left(1-\lambda_i^2\right)\right)}} \quad (5)$$

based on factor loadings. Both principal component and factor loadings are, essentially, PMCs between an item and a score variable θ (see Cramer & Howitt, 2004; Kim & Mueller, 1978; Yang, 2010), that is, $\lambda_i = \lambda_{i\theta}$ = PMC.

Because the estimates of item–score correlations are deflated, also the estimates of reliability are deflated.[1] In empirical settings with items with extreme difficulty levels, deflation in the estimates of reliability have been noted to be up to 0.60–0.70 units of reliability (see examples in, for instance, Gadermann et al., 2012; Metsämuuronen, 2022a, 2022c, 2022d; Metsämuuronen & Ukkola, 2019; Zumbo, Gadermann, & Zeisser, 2007). Hence, Zumbo and colleagues (2007; Gadermann et al., 2012) and Metsämuuronen (2022c, 2022d, 2022e) have offered different alternatives for the challenge of deflation. These options are called deflation-corrected estimators of reliability (DCER) and they are discussed later.

In the early days of test theory, the third concept, standard error of the measurement, was more important than the concept of reliability (see Gulliksen, 1950). However, by 1950, the concept on reliability superseded the concept of standard error when it comes to the interest in academic writings (see Gulliksen, 1950). However, these two concepts are closely linked because *S.E.m* is defined through reliability (see Eq. 1):

$$S.E.m = \sigma_E = \sigma_X\sqrt{1-REL} \quad (6)$$

(e.g., Gulliksen, 1950). It seems, however, that pendulum has swung back when it comes to weight of reliability and standard errors: in the large-scale testing settings such as PISA (Programme of International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study), instead of general reliability of the score, the interest is mainly in the standard errors in different parts of the ability scale (see, e.g., Foy & LaRoche, 2019; Schult & Sparfeldt, 2016). This has been motivated by two facts. First, reliability is usually a kind of *average* statistic related to the score and, hence, it seems not applicable in reflecting standard error in different parts of the ability scale (see discussion in, e.g., Metsämuuronen, 2022a, 2022b). Second, the testing settings related to several booklets with linked items, as is a standard procedure in the international settings and sometimes in the national level achievement testing, do not support using reliability in assessing the discrimination power of the combined score; we seem not have such estimator of reliability that would be generally accepted for these kinds of settings (see the discussion and possible options in Metsämuuronen, 2022b).

Deflation in the estimates of reliability binds together the concepts of correlation, reliability, and standard error of the measurement. On the one hand, primarily, the reason for the radical deflation in reliability is in the radical deflation in the ρ_{iX} or $\rho_{i\theta}$ embedded to most of the classical estimators of reliability of which magnitude is illustrated later. On the other hand, because of the deflation in the estimates of reliability, the estimates of standard error may be radically inflated. Metsämuuronen (2022a) gives examples of extreme real-life datasets where the standard errors are almost 20 times higher when using

[1] Notably, the underestimation in the estimates of reliability have been discussed widely in literature starting from Guttman (1945). Guttman showed that all his six estimators of reliability give underestimates. This generalizes to many other estimators such as coefficient alpha (see discussion and literature in Metsämuuronen, 2022d). The underestimation related to coefficient alpha has been discussed widely, and challenges related to alpha are well known although not necessarily well understood by general users (see discussions in, e.g., Cho & Kim, 2015; Hoekstra et al., 2019; Sijtsma, 2009). Therefore, there is an ongoing debate whether we should remove coefficient alpha from use (see the discussion in, e.g., Dunn, Baguley & Brunsden, 2013; McNeish, 2017; Sijtsma, 2009; Trizano-Hermosilla & Alvarado, 2016; Yang & Green, 2011) or not (see, e.g., Bentler, 2009; Falk & Savalei, 2011; Metsämuuronen, 2017; Raykov & Marcoulides, 2017; Raykov, West, & Traynor, 2014). This article does not discuss the traditional underestimation in estimators connected to modelling errors such as violations in tau-equivalency, unidimensionality, and uncorrelated errors. Deflation is a more profound source of underestimation, and it concerns not only coefficient alpha but also coefficients theta, omega, and rho even if the latter has often been suggested to replace alpha (see literature above).

traditional estimators of reliability in comparison with deflation-corrected estimators of reliability. In non-extreme datasets with widely deviating item difficulties, the standard errors based on traditional estimators of reliability may be two or three times higher in comparison with DCERs depending on which estimator of correlation and which estimator of reliability are in use (see Metsämuuronen, 2022b)—some examples of this phenomenon are given in the empirical section.

To condense the discussion above, traditional estimators of reliability such as coefficients alpha, theta, omega, and rho are prone to give deflated estimates of reliability for the test score when individual test items have extreme difficulty level in the target population. This is a typical pattern where the Pearson's point–biserial and point–polyserial coefficient of correlation between items and score (*Rit*) may be radically deflated causing radical deflation in the estimates of reliability and inflation in standard errors. The magnitude of deflation may be remarkable; 0.40–0.60 units of reliability have been reported in extreme cases.

These kinds of tests prone to produce deflated estimates of reliability are common in educational settings because tests are often structured to include both easy, medium, and demanding tasks. Exceptions of this logic may be tests that are aimed to measure a certain standard level, such as criterion-referenced licensure and certification examinations as is usual, among others, in language testing. Metsämuuronen (2018) discusses the matter and notes that the issues related to reliability are usually more highlighted in the norm-referenced testing than in the standards- or criterion-referenced testing. Reaching a certain standard level, that is, passing a test aimed to measure a certain level, does not depend on the great variability within the test takers or test items; even if *all* the test takers would pass the certain standard level, and the reliability would be zero because of technical reasons, the result may be acceptable. Then, the point made by Popham and Husek (1969, p. 3) makes sense; they noted that reliability indices based on variability in the dataset, as the traditional estimators of reliability usually are, "are *not only irrelevant to criterion-referenced uses but are actually injurious to their proper development and use*". However, Metsämuuronen (2018) points that, whenever the score or sub-scores are used as a basis for the *standard setting*, the reliability issues *are* relevant; if a score is used in the process, but the reliability of the score is very low, we cannot trust the score. Kane (1986, p. 221) suggests that "*the test-based procedure* [related to standard setting] *is found to improve the accuracy of universe score estimates only if the test reliability is above 0.50*". If a test used in the criterion- or standard-referenced testing includes wide variety of item difficulties, the issue of deflation in reliability is apparent.

With tests with wide variety of item difficulty, the deflation-corrected estimators (Metsämuuronen, 2022a, 2022c, 2022d, 2022e, 2022f) could be reasonable options; instead of *Rit*, they use better-behaving estimators of correlation as the linking factor between the item and the score variable; this matter is discussed further in Section "Conceptual differences between the traditional and deflation-corrected measurement models". Selecting wisely the linking coefficient, deflation-corrected estimators of reliability may produce significant advance in assessing the true reliability and true standard error related to the score, specifically, in the testing settings familiar in achievement testing.

Research Questions

This article discusses the effect of deflation in reliability from the viewpoint of achievement testing focusing, especially, on the characteristic form of tests in educational settings of including wide variety of item difficulties in the test. It is asked (and answered), first, *why* the estimates of reliability by the traditional estimators are, practically always, deflated with tests with wide variety of item difficulties and, hence, why standard errors of measurement are inflated when using the traditional estimators of reliability. Second, why the deflation-corrected estimators would be better estimators in achievement testing?

The traditional estimators are compared with selected deflation-corrected estimators of reliability. It is shown empirically that deflation-corrected estimators give estimates closer to the population value than the traditional estimators do. Hence, it is argued that the deflation-corrected estimators of reliability are reasonable alternatives for the traditional estimators, specifically, in the settings related achievement testing with tests with wide variety of item difficulties.

The course of the study starts with a brief conceptual discussion of the traditional and deflation-corrected measurement models after which empirical examples are given for the magnitude of deflation by the traditional estimators.

Conceptual differences between the traditional and deflation-corrected measurement models

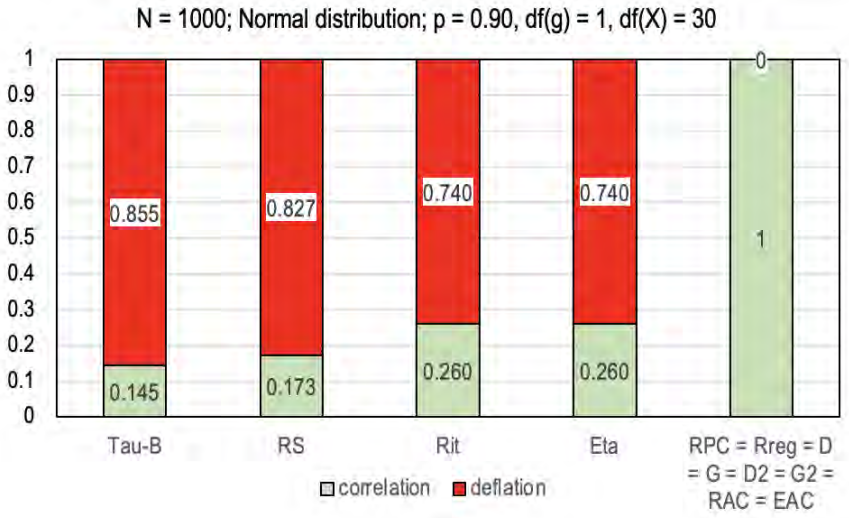
Reason for the deflation in reliability: item–score correlation

The root reason for the deflation in the estimates of reliability is in item–score correlation as discussed above. Metsämuuronen (2021a, 2022f) discusses in-depth the sources of deflation in item–score correlation. Based on simulations, seven sources of mechanical error in the estimates of correlation have been detected which all cause cumulative negative bias in PMC in general and in *Rit* specifically. According to simulations, PMC tends to underestimate the true association the more 1) the wider the number of categories in the variables differ from each other, 2) the more extreme is the item difficulty, 3) the further the distribution of the score is from the uniform distribution, 4) the less categories there are in the item, 5) the less categories there are in the score, 6) the less there are items forming the score because this has a strict connection to the number of categories in the scale of the score, and 7) the more tied cases (with non-uniform distribution) there are in the score. Although some of these sources are intertwined, they tend to be cumulative.

From the deflation viewpoint, the most challenging items are those that have extreme difficulty levels. On the one hand, it is good to include some very easy items in a test to give the lowest-performing test takers possibilities to show some achievement in the test. On the other hand, it is good to give possibilities also for the highest-performing test takers to show how far they can reach in their achievement and, hence, to include some demanding items in a test. This is the technical reason why achievement tests differ from the attitude tests; the achievement tests are often constructed by using items with widely deviating difficulty levels. Unfortunately, *Rit* is the most vulnerable with these extreme items. A simple example may illustrate the challenge.

Assume two identical (latent) variables with an obvious perfect correlation ($R = 1$). In practice, let us take a variable of $n = 1000$ normally distributed cases and double it. Of these identical variables with (obvious) perfect correlation, one (item *g*) is dichotomized into two categories (0–1) with difficulty level $p(g) = 0.90$ and the other (score X) is divided into 31 categories (0–30) with average difficulty level of $p(X) = 0.50$). The difference between the latent correlation and the observed correlation indicates strictly the magnitude of deflation in the estimates. Notably, such known estimators of correlation as *tau-b*

Figure 1. Deflation in selected estimators of reliability



(Kendall, 1948), Spearman rank-order correlation (R_S; Spearman, 1904), *Rit*, and coefficient *eta* (Pearson, 1903, 1905) cannot reach the latent perfect correlation but, instead, they include a remarkable magnitude of deflation (> 0.70 units of correlation) caused by technical and mechanical errors in the calculating process (Figure 1). On the contrary, such estimators as polychoric correlation coefficient (R_{PC}; Pearson, 1900, 1913), r-bireg and r-polyreg correlation (R_{REG}; Livingstone & Dorans, 2004, Moses, 2017, delta (D; Somers, 1962), gamma (G; Goodman & Kruskal, 1954), dimension-corrected D (D_2; Metsämuuronen, 2020b, 2021a), dimension-corrected G (G_2; Metsämuuronen, 2021a), attenuation-corrected *Rit* (R_{AC}; Metsämuuronen, 2022e, 2022g), and attenuation-corrected eta (E_{AC}; Metsämuuronen, 2022g) can detect the perfect correlation. Of the latter coefficients, D, D_2, and R_{REG} have minor defects in this matter; D and D_2 because of being affected by the tied cases and R_{REG} because of being affected by short tests (see Metsämuuronen, 2022f). The consequence of the deflation in *Rit* to the measurement model, reliability, and standard errors is discussed in what follows.

Measurement model including the element of deflation

The traditional measurement models do not include elements related to deflation; they assume that the measurement is deflation-free which is a too optimistic assumption. Metsämuuronen (e.g., 2022a, 2022d, 2022f) uses the term "mechanical error in the estimates of correlation" (MEC) to conceptualize the phenomenon. The general, simplified, one-latent variable measurement model combining the latent variable (θ), the observed values of an item g_i (x_i), and a weight factor w_i that links θ with x_i can be expressed as follows:

$$x_i = w_i\theta + e_i, \tag{7}$$

(e.g., Metsämuuronen, 2022a, 2022c) generalized from the traditional model (e.g., Cheng et al., 2012; McDonald, 1999) where e_i refers to the measurement error. The latent variable θ may be manifested in different forms as a compilation of the test items: raw score, principal component score, factor score, IRT-score, or various non-linear combinations of the items. Also, in the general model, the weight factor may vary although usually it is item–score correlation in some form including the estimators discussed above (*Rit*, R_{PC}, R_{REG}, D, D_2, G, G_2, R_{AC} and E_{AC}, or principal component or factor loading λ_i).

While knowing that a certain part of the measurement error is strictly technical or mechanical in nature (see Figure 1), but its magnitude could be reduced by selecting wisely the weight factor, Metsämuuronen (2022c, 2022d, 2022e) suggests to reconceptualize the classic relation of $X = T + E$ as $X = T + (E_{Random} + E_{MEC})$, where the new element related to deflation, E_{MEC}, is visible. Consequently, the measurement model in Eq. (7) can be reconceptualized as

$$x_i = w_i \times \theta + \left(e_{i\_Random} + e_{wi\theta\_MEC}\right), \tag{8}$$

where the element $e_{wi\theta\_MEC}$ refers to the fact that the magnitude of the mechanical error in the model depends on the weighting factor w_i, characteristics of the item i, and the manifestation of the score variable θ. Notably, the magnitude of the error in the models Eq. (7) and Eq. (8), that is, e_i and $\left(e_{i\_Random} + e_{wi\theta\_MEC}\right)$ respectively, is equal but the element related to deflation is visible in the latter form. If we select the weight element wisely so that the deflation is zero or near-zero, the element related to the technical or mechanical error approximates zero, that is, $e_{wi\theta\_MEC} \approx 0$. Then, if we use estimators of correlation that are deflation-free or close, we get MEC- or deflation-corrected (DC) measurement model:

$$\begin{aligned} x_i &= w_{i\_DC} \times \theta + \left(e_{i\_Random} + e_{wi\theta\_MEC}\right) \\ &= w_{i\_DC} \times \theta + e_{i\_DC} \\ &\approx w_{i\_DC} \times \theta + e_{i\_Random} \end{aligned} \tag{9}$$

Because the coefficient of correlation gives identical results with original variables and standardized versions of the variables, we can assume that both the items and score are standardized with the mean of 0 and variance 1. Then, this conceptualization leads to item-wise deflation-corrected error variance ($\Psi^2_{i\_DC}$):

$$Var\left(e_{i\_DC}\right) = \psi^2_{i\_DC} = 1 - w^2_{i\_DC}, \tag{10}$$

Assuming that the errors do not correlate, this measurement model generalizes to the compilation of items as follows:

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} w_{i\_DC} \times \theta + \sum_{i=1}^{k} e_{i\_DC}, \tag{11}$$

where $e_{i\_DC} \sim N(0, \Psi^2_{i_{DC}})$. The deflation-corrected error variance related to the test score can be written as

$$Var\left(\sum_{i=1}^{k} e_{i\_DC}\right) = \sum_{i=1}^{k} \psi^2_{i\_DC} = \sum_{i=1}^{k}\left(1 - w^2_{i\_DC}\right) \quad . \tag{12}$$

This conceptualization leads to short-cuts to deflation-corrected estimators of reliability.

Deflation-corrected estimators of reliability

The traditional estimators of reliability in Eqs. (2) to (5) assume that the measurement is deflation-free as noted above. A small modification gives the possibility to use them as bases for the deflation-corrected estimators. Theoretical bases for different families of DCERs discussed by Metsämuuronen (e.g., 2022c, 2022d, 2022e) are either based on alpha (Eq. 2):

$$\rho_{\alpha\_wi\theta} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times w_{i\theta}\right)^2}\right), \tag{13}$$

theta (Eq. 3):

$$\rho_{TH\_wi\theta} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} w^2_{i\theta}}\right), \tag{14}$$

omega (Eq. 6):

$$\rho_{\omega\_wi\theta} = \frac{\left(\sum_{i=1}^{k} w_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} w_{i\theta}\right)^2 + \sum_{g=1}^{k}\left(1 - w^2_{i\theta}\right)}, \tag{15}$$

or rho (Eq. 7):

$$\rho_{MAX\_wi\theta} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k}\left(w^2_{i\theta}/\left(1 - w^2_{i\theta}\right)\right)}}, \tag{16}$$

although other bases could be used as well. Basically, the element $w_{i\theta}$ refers to a fact that the magnitude of the estimate depends on three things: the characteristics of the weight factor (w), of the item (i), and of the score variable (θ) as a manifestation of the latent trait as discussed above. If $w_{i\theta}$ in Eq. (13) is operationalized as Rit, we get the traditional coefficient alpha. If, however, $w_{i\theta}$ is operationalized, for instance, as Somers D, we get a conservative option for the deflation-corrected estimator of reliability based on alpha. If in Eq. (14) $w_{i\theta}$ is operationalized as the principal component loading related to the first or the only principal component, we get the traditional coefficient theta. If, however in Eq. (14), $w_{i\theta}$ is operationalized as D, we get a conservative option for the deflation-corrected theta. Parallel, if in Eqs. (15) and (16) $w_{i\theta}$ is operationalized as factor loadings, we get the traditional omega and rho. By using D instead of the traditional factor loading, we get a conservative option for the deflation-corrected omega and rho.

Metsämuuronen (2022d) has typologized DCERs. Using R_{PC} and R_{REG} as $w_{i\theta}$ leads to theoretical reliability because they refer to inferred correlation between non-observed variables (see Chalmers, 2017). Using the other well-behaving estimators (D, G, D_2, G_2, RAC, or EAC) leads to more practical interpretations of the reliability. For example, the use of D or G gives the interpretation of indicating the proportion of logically (incrementally) ordered observations in all items on average after they are ordered by the score (see Metsämuuronen, 2021b, 2022d, 2023a, 2023b).

DCERs can be divided into two families: in *attenuation-corrected* estimators of reliability (ACER), $w_{i\theta}$ is operationalized as attenuation-corrected estimators of correlation (R_{AC} or E_{AC}; Metsämuuronen, 2022e, 2022g) and, in *MEC-corrected* estimators (MCER), $w_{i\theta}$ is replaced by totally *different* estimator of correlation such as D or G (see Metsämuuronen, 2022c, 2022d). However, here, both are called by a common name, deflation-corrected estimators. Notably, Zumbo's and colleagues (2007; Gadermann et al., 2012) ordinal alpha and ordinal theta may be included also in the extended family of DCERs; instead of changing the linking factor, the matrix of PMC is replaced by a matrix of R_{PC}.

Finally, it may be wise to take seriously the note by Metsämuuronen (e.g., 2022c, 2022e) that using theta, omega, and rho outside of their traditional context is debatable. In the framework of DCERs, it is assumed that these estimators *could* be used as independent estimators; this seems consistent with the general measurement model discussed above. Alternatively, it is possible to think that the estimators using R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, or E_{AC} instead of the traditional λ_i

are based on renewed procedures on principal component- and factor analysis where the factor loadings are, for example, R_{PC} and G_2 instead of PMC (cl. ordinal theta by Zumbo and colleagues, 2007).

Empirical examples of deflation of reliability in the achievement testing

Three examples are given of the phenomenon of deflation, specifically, within the achievement testing. The first example is a hypothetical example which is used in showing the manual calculation of the estimates. Two others are based on real-world datasets from national level testing settings. One is a dataset referred to above (Metsämuuronen & Ukkola, 2019) with extremely easy test items; this was a screening test of language related to the administrative language of the test itself. Only the students with immigrant background were expected to make mistakes in the test. Consequently, 75% of the students got full marks in the 8-item, 11-points test. The other dataset is based on a 30-item test of mathematics which is used as basis for a simulation of the performance of the estimators. From the original dataset, 1,440 tests with different sample sizes were produced. This dataset is used to study the behaviour of the estimators in comparison with the "population". This analysis is intensified by a smaller dataset with 560 short tests with more extreme item difficulties from the same basic dataset.

In the numerical examples, of the alternative estimators of correlation, *D*, *G* and R_{AC} are used for binary cases and D_2, G_2 and R_{AC} for polytomous cases. The outcomes are similar type with other better-behaving estimators (see Metsämuuronen, 2022d). *D* represents a conservative estimator and *G* more liberal estimator of correlation. R_{AC} represents attenuation-corrected alternative, while *D* and *G* lead to MEC-corrected alternative for DCERs.

Case 1: A hypothetical dataset with widely deviating item difficulties

The first case is a small dataset ($n = 12$, $k = 5$) with incremental difficulty levels in items ($p = 0.083$–0.917). This could be a short subtest of "Sets" or "Programming" amid a larger test battery related to mathematics achievement. Relevant indicators of the items such as item variances, maximal correlation in the dataset, as well as indices of item–score correlation (*Rit*, *D*, *G*, and R_{AC}) and relevant derivatives of the estimators of correlation for estimating the reliability are collected in Table 1a. Calculation of the estimators of correlation is discussed in Appendix 1. Table 1b shows the principal component and factor loadings needed for the traditional theta, omega, and rho. Table 1c collects the estimates of reliability.

From Table 1a it is known that *Rit* varies 0.302–0.704 although it is also known that item g1, as an example, cannot even get higher value than what already was obtained, $Rit = Rit^{max} = 0.427$ and, hence, attenuation-corrected *Rit* equals $R_{AC} = Rit/Rit^{max} = 1$.[2] Notably, with items with extreme difficulty levels, the estimates by *D*, *G*, and R_{AC} are remarkably higher (0.909–1.000) in comparison with *Rit* (0.472). The deflation-corrected estimators of correlation can detect the deterministic pattern by the high magnitude in the estimates of association: $D = G = RAC = 1$. In the case of item g1, $D = 0.909$ detects the one tied pair while $G = 1$ ignores it. Notably, in the binary case, *D* equals D_2, *G* equals G_2 and *Rit* equals coefficient *eta*.

Using Eq. (2) and Table 1a, the estimate by the traditional alpha is $\hat{\rho}_\alpha = \frac{5}{4}\left(1 - \frac{0.778}{0.957^2}\right) = 0.189$. Parallel, using Eq. (3) and Table 1b, the traditional estimate by theta is $\hat{\rho}_{TH} = \frac{5}{4}(1 - 1/1.598) = 0.486$, traditional omega by Eq. (4) and Table 1b is $\hat{\rho}_\omega = 1.485^2/(1.485 + 3.600) = 0.280$, and the traditional rho by Eq. (5) and Table 1b is $\hat{\rho}_{MAX} = \frac{1}{1+\frac{1}{499.250}} = 0.998$. The last seems and feels obvious overestimation caused by item g5 which behaves poorly in the maximum likelihood estimation and causes a near-deterministic pattern (see Table 1b). It is good to remember the warnings by Aquirre-Urreta and colleagues (2019) and Metsämuuronen (2022a) that maximal reliability easily gives (obvious or suspicious) overestimates with finite samples. In small samples the

[2] The maximum value of *Rit* and *eta* is obtained by ordering the items and score independently and calculating the correlation after that (Metsämuuronen, 2022e, 2022g).

Table 1a. Hypothetic dataset with widely deviating item difficulties

	items					scores		
test taker	g1	g2	g3	g4	g5	θ_X	θ_{PC}	θ_{FA}
1	1	0	0	0	0	1	-0.168	-0.289
2	0	1	0	0	0	1	-2.008	-0.289
3	1	1	0	0	0	2	-0.637	-0.288
4	1	0	1	0	0	2	0.368	-0.289
5	1	1	0	0	0	2	-0.637	-0.288
6	1	1	0	0	0	2	-0.637	-0.288
7	1	0	1	1	0	3	1.487	-0.288
8	1	1	1	0	0	3	-0.101	-0.289
9	1	1	1	0	0	3	-0.101	-0.289
10	1	1	1	0	0	3	-0.101	-0.289
11	1	1	0	1	1	4	1.520	3.174
12	1	1	1	1	0	4	1.018	-0.287
p	0.917	0.750	0.500	0.250	0.083	SUM		
σ^2_i	0.076	0.188	0.250	0.188	0.076	0.778		
Rit^{max}	0.472	0.689	0.912	0.770	0.487			
Rit	0.472	0.302	0.522	0.704	0.472	2.472		
$D = D2$	0.909	0.370	0.611	0.889	0.909	3.689		
$G = G2$	1	0.500	0.688	1	1	4.188		
$RAC = Rit/Rit^{max}$	1	0.438	0.572	0.914	0.971	3.894		
σ_i x Rit	0.131	0.131	0.261	0.305	0.131	0.957		
σ_i x D	0.251	0.160	0.306	0.385	0.251	1.353		
σ_i x G	0.276	0.217	0.344	0.433	0.276	1.546		
σ_i x RAC	0.276	0.190	0.286	0.396	0.268	1.416		
D^2	0.826	0.137	0.373	0.790	0.826	2.954		
$1-D^2$	0.174	0.863	0.627	0.210	0.174	2.046		
$D^2/(1-D^2)$	4.762	0.159	0.596	3.765	4.762	14.044		
G^2	1	0.250	0.473	1	1	3.723		
$1-G^2$	0	0.750	0.527	0	0	1.277		
RAC^2	1	0.192	0.328	0.835	0.942	3.296		
$1-RAC^2$	0	0.808	0.672	0.165	0.058	1.704		

Table 1b. Principal component and factor loading

	PC loadings		Factor loadings			
	λ_i	λ_i^2	λ_i	λ_i^2	$1-\lambda_i^2$	$\lambda_i^2/(1-\lambda_i^2)$
g1	0.632	0.399	0.091	0.008	0.992	0.008
g2	-0.339	0.115	0.174	0.030	0.970	0.031
g3	0.447	0.200	-0.301	0.091	0.909	0.100
g4	0.809	0.654	0.522	0.272	0.728	0.375
g5	0.479	0.229	0.999	0.998	0.002	499.250
SUM		1.598	1.485	1.400	3.600	499.764

Table 1c. Estimates of reliability

	base			
weight	Alpha	Theta	Omega	Rho
traditional	0.189	0.468	0.280	0.998
D	0.719	0.827	0.869	0.934
G	0.843	0.914	0.932	(no solution)
RAC	0.765	0.871	0.899	(no solution)

probability to obtain deterministic of a near-deterministic datasets at least in one item is very high. In these cases, the estimates by rho will be obviously too high. Except the obvious overestimate by rho, the estimates by the traditional magnitudes of the estimates are too low to be accepted as real estimates. The better behavior of the estimators of correlation by *D*, *G*, and R_{AC} with items with extreme difficulty level gives a hint that the traditional estimators are radically deflated.

Because the magnitude of the estimates by *D*, *G*, and R_{AC} tends to be higher than those by *Rit*, that is, they give deflation-corrected estimates of correlation, also the magnitude of the estimates of reliability by DCERs are higher than by using the traditional estimators. The deflation-corrected estimates are calculated by using the estimators in Eq. (13) to (16) and Table 1a. Then, the raw score is used as the manifestation of the latent ability—using the principal component-, factor- or IRT-score would not change the results much (see Metsämuuronen, 2022c). Here, the deflation-corrected estimates are computed by using *D* as an example; the others are computed in parallel manner. By using Eq. (13) and Table 1a, deflation-corrected alpha using *D* as the linking factor, "*alphaD*", is $\hat{\rho}_{\alpha\_D} = \frac{5}{4}\left(1 - 0.778/1.353^2\right) = 0.719$. Using Eq. (14) and Table 1a, the deflation-corrected theta, "*thetaD*" is $\hat{\rho}_{TH\_D} = \frac{5}{4}\left(1 - 1/2.954\right) = 0.827$. The deflation-corrected omega, "*omegaD*" is calculated by Eq. (15) and Table 1a as follows: $\hat{\rho}_{\omega\_D} = 3.689^2 / \left(3.689^2 + 2.046\right) = 0.869$. Unlike with *G* and R_{AC}, it is possible to also compute the deflation-corrected rho by using *D* as the weight; using *G* or R_{AC} would not be possible because of the deterministic pattern in one or several items. Then, the deflation-corrected rho by using *D* as the linking factor, "*rhoD*", is calculated by Eq. (16) and Table 1a as follows: $\hat{\rho}_{MAX\_D} = \left(1 + \frac{1}{14.044}\right)^{-1} = 0.934$. The magnitude of the last estimate seems not to refer to an obvious overestimate unlike by using the traditional rho. However, Metsämuuronen (2022d) do not suggest using DCERs based on rho with small sample sizes ($n < 200$) because of the risk of obvious overestimation or because the estimate is not possible to be calculated because of deterministic patterns.

The DCERs using *G* give higher estimates than those by *D*. This is expected because the estimates by *G* are almost always higher than those by *D* (see exceptions in Metsämuuronen, 2021b). This is caused by different base for the probability; *G* omits the tied cases while *D* use all cases. The behavior of R_{AC} in comparison with *D* and *G* is largely unstudied. However, it seems that estimates are systematically somewhat higher in magnitude than those by *D* but lower than those by *G*. Hence, if the estimates using *D* are more conservative and those by *G* are more liberal, the estimates using R_{AC} seem to be a kind of consensus between liberal and conservative estimates. If we take *alphaD*, *thetaD*, and *omegaD* as reference, in the given dataset, the traditional alpha is deflated by 74% [=(0.719−0.189)/0.719×100%], theta by 43% and omega 68%. If the estimators using *G* were the reference, the traditional alpha was deflated by 77%, theta by 49% and omega by 69%.

Deflation in the estimates of reliability is seen also as inflation in the standard errors. Knowing that the standard deviation of the score is $\sigma_X = 0.957$, it is known that the traditional standard error would be *S.E.m.* $= 0.957 \times \sqrt{1 - 0.189} = 0.862$ points if coefficient alpha is used in estimation. If DCERs are used, the standard error would be between 0.379 (*G*) and 0.507 (*D*) points. This means a reduction of 41–56% in the estimate of standard error if DCERs are used instead of the traditional estimators of reliability. In other words, using the traditional estimator of

reliability gives 1.7–3.2 times wider standard errors in comparison with estimates based on DCERs. The difference would be wider (2.3–3.3 times) if omega would be used as a base in estimation.

Case 2: Test of extreme difficulty level

Sometimes, the achievement test may be very (or too) easy or difficult for the target group—or it is purposefully made easy for a specific reason such as being a diagnostic part of a larger test set. Sometime, this kind of (sub)test may be aimed to be a test for selecting test takers to continue at a specific level in the rest of the test—this is common in adaptive testing—or as a test in a low-level criterion-referenced standard. Because all items in the dataset may be very easy to the target population, the traditional estimators are expected to be radically deflated. As an example of this kind of test, a specific, national level dataset with exceptionally easy items with n = 7,770 test takers discussed and further analyzed by Metsämuuronen (2022a, 2022d, 2022f; originally in Metsämuuronen and Ukkola, 2019) is used here as an example.

The test was a screening test of proficiency in the language used in the factual test; only the test-takers with immigrant background with assumingly lower achievement level in the administrative language of the school (and in the test) were expected to make mistakes in the test items. Altogether 75% of the test takers gave full marks in the test of 8-items and 11-points; the distribution of the score is obviously non-normal (Figure 2). The question is, can the traditional estimators of reliability detect the fact that the lowest-performing test takers are systematically scoring lower also in the individual items? Descriptive statistics of the dataset are discussed by Metsämuuronen (2022d). Here, relevant pieces of information for estimating reliability are collected in Table 2a, principal component- and factor loadings for the traditional theta, omega, and rho in Table 2b, and the estimates of reliability in Table 2c. The factual calculation of estimates is demonstrated in Case 1 and, hence, only the outcome is of interest here.

The traditional estimators of reliability are obviously deflated; estimates vary 0.246 (alpha) to 0.493 (rho) indicating, traditionally, that the test cannot separate between the lower- and higher performing test takers. However, the average item–score associations by D and G are 0.831 and 0.879, respectively, indicating high association between the score and individual items. Knowing the interpretation of D and G, on average, 92–94% of the observations are logically (incrementally) ordered to the same order as the score is, that is, almost all lower-performing test takers perform lower in *all* items of the test after they are ordered by the score (see the discussion of common language estimators of reliability in Metsämuuronen, 2023a). This is known from the fact that the common language effect sizes $PHD = 0.5+0.5\times0.831 = 0.915$ if the tied pairs are included (D) and $PHG = 0.5+0.5\times0.879 = 0.940$ if the tied pairs are omitted (G) (see the discussion of common language estimators of effect size in Metsämuuronen, 2023b). The estimates of reliability by using D and G as the linking factors

Figure 2. Distribution of the original dataset in Case 2 (n = 7,770)

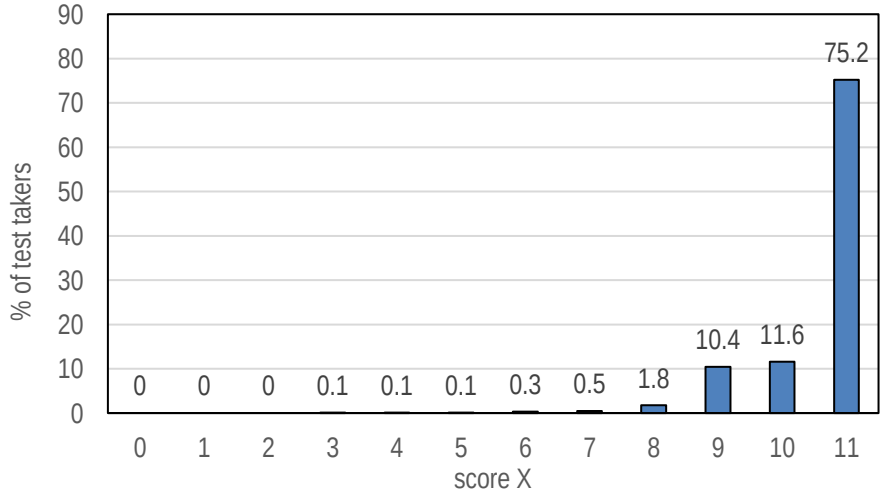


Table 2a. Characteristics of items, estimates of item–score correlation, and derivatives for estimating reliability

	g1	g2	g3	g4	g5	g6	g7	g8	SUM
scale	0−1	0−1	0−1	0−1	0−2	0−1	0−2	0−2	
p	0.960	0.980	0.990	0.910	0.890	0.980	0.985	0.990	
σ_i	0.186	0.126	0.088	0.287	0.610	0.122	0.211	0.169	
σ^2_i	0.035	0.016	0.008	0.082	0.372	0.015	0.045	0.028	0.600
Rit^{max}	0.635	0.548	0.478	0.754	0.811	0.541	0.579	0.553	
Rit	0.350	0.268	0.288	0.455	0.747	0.258	0.329	0.376	3.071
D	0.791	0.779	0.858	0.789	0.952	0.766	0.832	0.877	6.644
G	0.857	0.846	0.911	0.789	0.979	0.831	0.897	0.924	7.033
$RAC = Rit/Rit^{max}$	0.551	0.490	0.603	0.603	0.921	0.477	0.568	0.680	4.892
σ_i x Rit	0.065	0.034	0.025	0.131	0.455	0.032	0.069	0.063	0.875
σ_i x D	0.147	0.098	0.076	0.226	0.581	0.094	0.176	0.148	1.546
σ_i x G	0.160	0.107	0.080	0.226	0.597	0.102	0.189	0.156	1.617
σ_i x RAC	0.103	0.062	0.053	0.173	0.562	0.058	0.120	0.115	1.245
D^2	0.625	0.607	0.736	0.622	0.907	0.586	0.692	0.770	5.546
$1-D^2$	0.375	0.393	0.264	0.378	0.093	0.414	0.308	0.230	2.454
$D^2/(1-D^2)$	1.670	1.546	2.789	1.647	9.768	1.417	2.242	3.345	24.425
G^2	0.734	0.716	0.829	0.622	0.958	0.691	0.804	0.854	6.209
$1-G^2$	0.266	0.284	0.171	0.378	0.042	0.309	0.196	0.146	1.791
$G^2/(1-G^2)$	2.760	2.520	4.856	1.647	22.828	2.236	4.105	5.871	46.823
RAC^2	0.304	0.240	0.363	0.364	0.849	0.228	0.322	0.462	3.131
$1-RAC^2$	0.696	0.760	0.637	0.636	0.151	0.772	0.678	0.538	4.869
$RAC^2/(1-RAC^2)$	0.436	0.315	0.570	0.571	5.619	0.295	0.475	0.859	9.141

Table 2b. Principal component and factor loading

	PC loadings		Factor loadings			
	λ_i	λ_i^2	λ_i	λ_i^2	$1-\lambda_i^2$	$\lambda_i^2/(1-\lambda_i^2)$
g1	0.447	0.200	0.276	0.076	0.924	0.082
g2	0.430	0.185	0.260	0.068	0.932	0.073
g3	0.605	0.366	0.471	0.222	0.778	0.285
g4	0.468	0.219	0.291	0.085	0.915	0.093
g5	0.204	0.042	0.111	0.012	0.988	0.012
g6	0.375	0.141	0.213	0.045	0.955	0.048
g7	0.288	0.083	0.160	0.026	0.974	0.026
g8	0.633	0.401	0.512	0.262	0.738	0.355
SUM		1.636	2.294		7.204	0.974

Table 2c. Estimates of reliability

	base			
weight	Alpha	Theta	Omega	Rho
traditional	0.246	0.444	0.422	0.493
D	0.856	0.937	0.947	0.961
G	0.880	0.959	0.965	0.979
RAC	0.700	0.778	0.831	0.901

correspond well with this fact; estimates using *D* vary 0.856 (*alphaD*) to 0.961 (*rhoD*) and the estimates using *G* vary 0.880 (*alphaG*) to 0.979 (*rhoG*). Hence, if using the estimators with *D* as the weight factor, the traditional estimators are deflated by 71% (alpha), 53 % (theta), 55% (omega), and 49% (rho). Correspondingly, if the estimators using *G* are used as the reference, the deflation rates are 72%, 54%, 56%, and 50%, respectively.

The deflation of this size in the estimates of reliability have obvious consequences to the estimates of standard error. While knowing that the population standard deviation of the score variable *X* is $\sigma_X = 0.875$, it is known that the average standard error would be *S.E.m.* = $0.875 \times \sqrt{1 - 0.246} = 0.760$ points if the traditional coefficient alpha is used in estimation and 0.622 points if rho is used. If DCERs are used, the standard error would be between 0.127 (*rhoG*) and 0.332 points (*alphaD*). This means a reduction of 56–80% in the estimate of standard error if DCERs are used instead of the traditional estimators of reliability. In other words, in the case, using the traditional estimator of reliability gives 2.3–4.9 times wider standard errors in comparison with estimates based on DCERs.

Case 3: Larger simulation of the behaviour of DCERs with a special interest on the tests with items with extreme difficulty levels

Datasets and variables. Cases 1 and 2 are specific extreme cases where the population estimates are not known. Case 3 presents a simulation based on a real-life dataset with an artificial "population" that can be used in assessing the characteristics of the estimators from the viewpoint of their behavior with population estimates. A dataset of 4,023 nationally represented test-takers of achievement in mathematics with 30 binary items (FINEEC, 2018) is used as the "population". In the original dataset, $\rho_a = 0.885$, $\rho_{TH} = 0.890$, $\rho_\omega = 0.887$, and $\rho_{MAX} = 0.895$, the traditional item–score correlation range $0.332 < Rit < 0.627$ with the average $\overline{Rit} = 0.481$, and item difficulties range $0.24 < p < 0.95$ with the average $\bar{p} = 0.63$.

Of the original "population", 40 smaller samples with finite or small sample sizes were drawn, and 1,440 tests were formed by varying the number of test-takers (n = 25, 50, 100, and 200), test items (k = 2–30, $\bar{k} = 10.33$), categories in the items ($df(g)$ = 1–15, $df(g) = 4.57$) and in the score ($df(X)$ = 10–27, $\overline{df(X)} = 18.06$), and the average difficulty levels in items ($\bar{p}$ = 0.50–0.76, $\bar{\bar{p}}$= 0.66). Consequently, the lower bound of reliabilities vary notably (ρ_a = 0.55–0.93, $\overline{\rho_\alpha}$). The polytomous items were formed by summing the binary items. Hence, the datasets are partly dependent, specifically, when it comes to the score variable: the dataset is formed so that the same score is common to several combination of items. This has an effect to the standard errors: both the binary and polytomous items lead to the same standard error because the score variance is common for both types of items. The dataset of individual items (n = 14,880) including relevant indicators of item–score association is available in CSV format at http://dx.doi.org/10.13140/RG.2.2.10530.76482 and in SPSS format at http://dx.doi.org/10.13140/RG.2.2.17594.72641. The dataset of reliabilities (n = 1,440) is available in CSV format at http://dx.doi.org/10.13140/RG.2.2.30493.03040 and in SPSS format at http://dx.doi.org/10.13140/RG.2.2.27971.94241.

As the main utility of DCERs comes with tests including items with wide variety of difficulty levels or test with extreme in difficulty level, an additional dataset with shorter tests with more extreme difficulty levels was prepared from the original "population". From the population, ten random samples with sample sizes n = 200, n = 100, n = 50, n = 25 were formed as in the main dataset. However, in these 40 samples, 14 shorter tests with k = 5 (n = 4), k =8 (n = 3), and k = 10 (n = 7) binary items were formed. Selecting shorter compilations of items made it possible to select the most extreme items to the test. Of the 14 tests, five used the easiest items with different compilations, five used the most difficult items, and four used the items of medium difficulty levels. In this additional dataset with 560 tests, the traditional coefficient alpha (*AlphaRit*) and rho (*Maxrel*) were computed and the conservative *AlphaD* is used as the benchmark. This dataset is available in CSV format at http://dx.doi.org/10.13140/RG.2.2.18911.94887 and in SPSS format at http://dx.doi.org/10.13140/RG.2.2.25622.83521.

In what follows, coefficient alpha is mainly used as an example. This may be justified because it is the most

widely used estimator of (the lower bound of) reliability (see, literature in, e.g., Cho & Kim, 2015; Hoekstra et al., 2019; Sijtsma, 2009). An interested reader finds relevant comparisons using omega as an example in Metsämuuronen (2022a) and (2022d). Some points are lifted related to the capability of traditional and deflation-corrected estimators to reflect the population reliability. A simple statistic is used for this: the difference between the sample estimate and the population value (*d*). When $d < 0$, the sample estimate underestimates the population value and when $d > 0$, reliability is overestimated. In what follows, this deviation between the sample estimate and population value related to a specific estimator is referred to as "*dAlphaD*" or "*dAlpha* traditional". Because the real population reliability is unknown as being a real-world dataset, each estimator runs its own race: *AlphaD* in the sample is compared with *AlphaD* in the population, as an example.

General comparison of traditional and deflation-corrected estimators. The first, general lift from the datasets is that, in the binary settings, the DCERs give a univocal message that the reliability is at the level 0.914–0.938 rather than 0.846–0.884 as suggested by the traditional estimators (Table 3, Figure 3). Within each base, the deviance obtained by different weight factors is nominal (0.914–0.926 by alpha; 0.926–0.935 by theta; 0.929–0.938 by omega; 0.944–0.955 by rho). However, Figure 3 illustrates the fact that the estimates based on rho may be mildly overestimated. This is expected in the datasets with finite samples (see Aquirre-Urrata et al., 2019; Metsämuuronen, 2022a).

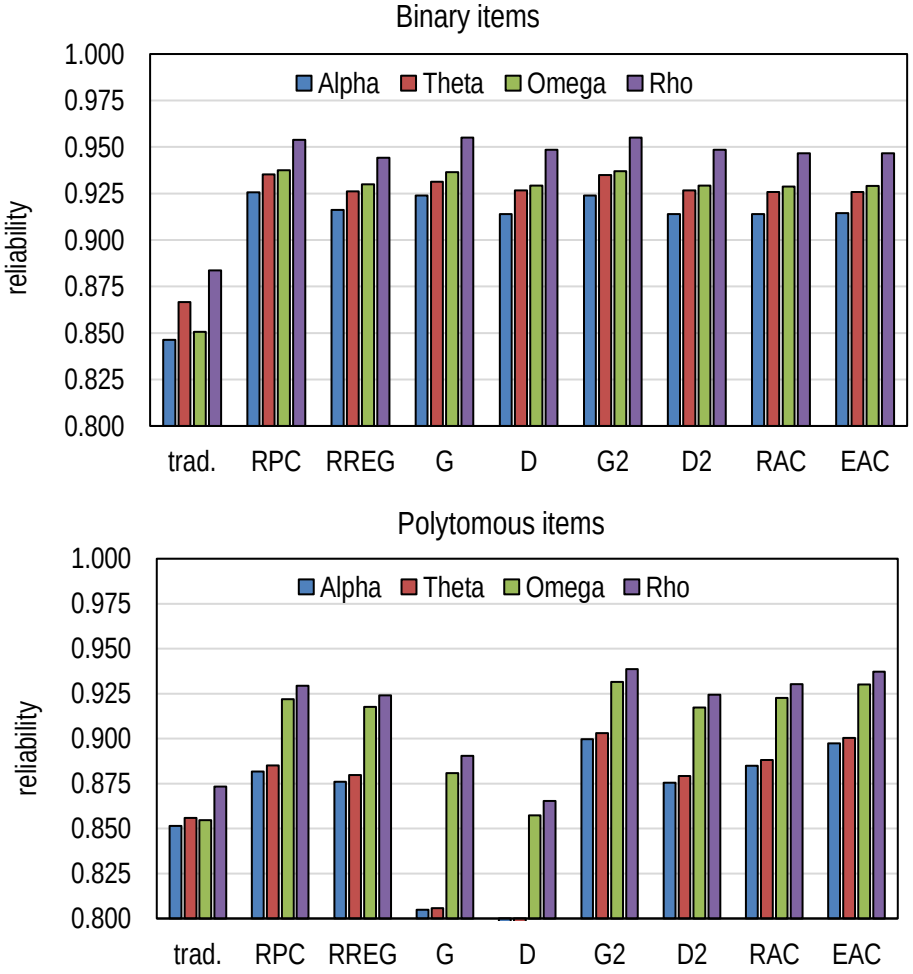
With polytomous items, it is obvious that the estimators using *D* and *G* as the weighting factor fail to reach the real reliability (Table 3, Figure 3). This is caused by their poor behavior in reflecting the item–score association when the number of categories exceeds 3 (in *D*) or 4 (in *G*) (see Metsämuuronen, 2021a, 2021b). In the polytomous case, DCERs based on omega and rho seem to be close to each other, and their magnitude is notably higher than those based on alpha and theta. The former suggest that the reliability would be at the level of 0.917–0.937 and the latter that the level would be at 0.876–0.903. Both are somewhat higher than what the traditional estimators suggest (0.851–0.873). Deflation in the dataset is mild in comparison with the extreme datasets in Cases 1 and 2; in binary case, on average 7–9% and, in polytomous case, 4–7% depending on the base and the weight factor.

Second, on average, of the traditional estimators, alpha, theta, and omega are conservative: they tend to give estimates that are lower in magnitude than the corresponding population value is. Rho is liberal: with small sample sizes, it tends to give estimates that overestimate the population value as discussed above. Although each estimator produces sample estimates that are higher or lower than the population value, rho

Table 3. Means of estimates of reliability in the simulation dataset (n = 1,440 estimates)

	binary items					polytomous items				
	N	Alpha	Theta	Omega	Rho	N	Alpha	Theta	Omega	Rho
traditional	320	0.846	0.867	0.851	0.884	1120	0.851	0.856	0.855	0.873
RPC	320	0.926	0.935	0.938	0.954	1120	0.882	0.885	0.922	0.929
RREG	320	0.916	0.926	0.930	0.944	1120	0.876	0.880	0.918	0.924
G	320	0.924	0.931	0.936	0.955	1120	0.805	0.806	0.881	0.890
D	320	0.914	0.927	0.929	0.949	1120	0.753	0.758	0.857	0.865
G2	320	0.924	0.935	0.937	0.955	1120	0.900	0.903	0.932	0.939
D2	320	0.914	0.927	0.929	0.949	1120	0.876	0.879	0.917	0.924
RAC	320	0.914	0.926	0.929	0.947	1120	0.885	0.888	0.923	0.930
EAC	320	0.914	0.926	0.929	0.947	1120	0.897	0.900	0.930	0.937

Figure 3. Average estimates in binary and polytomous settings

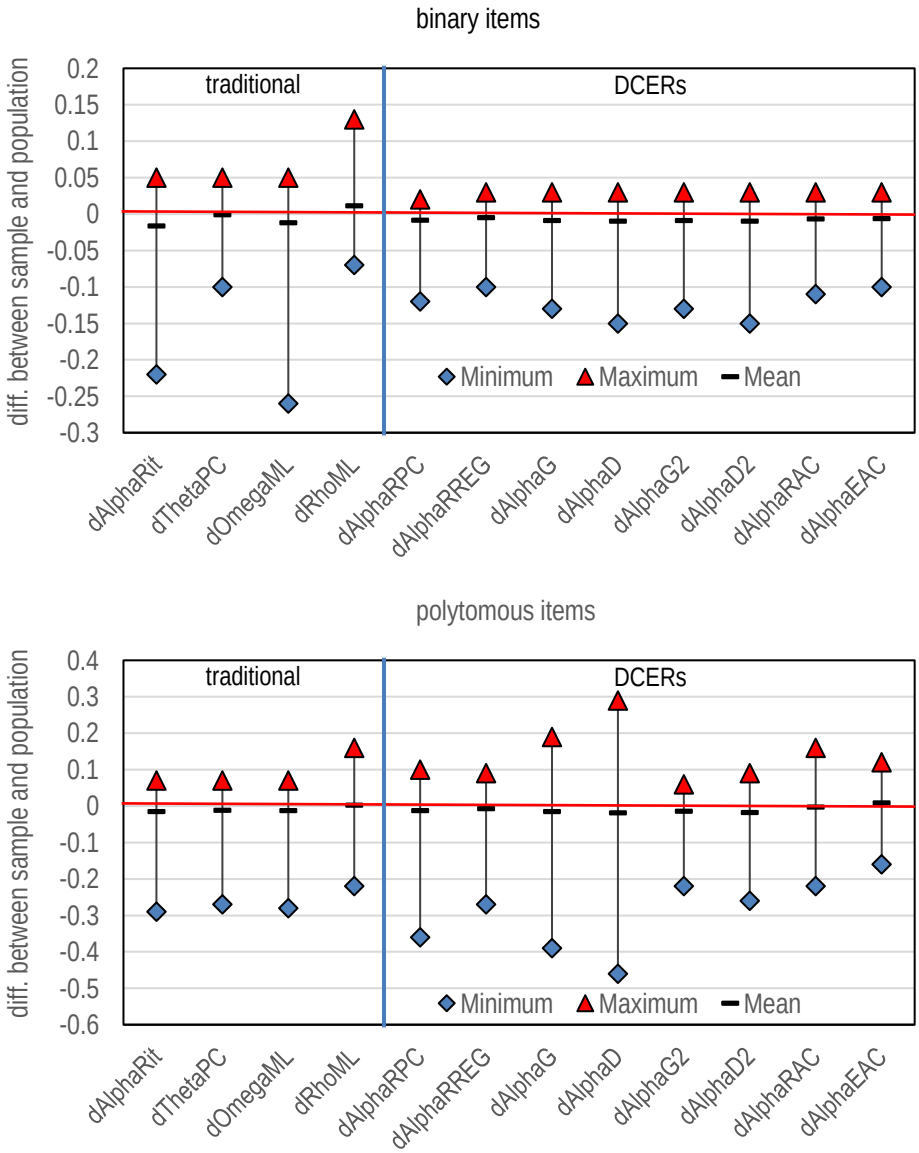


is prone to produce more overestimates than underestimates. This is specifically true with binary items (Figure 4). Notably, in the binary case, all DCERs are conservative but the estimates are closer to the population value in comparison with the traditional estimator when it comes to range. This is specifically true with alpha, omega, and rho. With polytomous items, it is not recommendable to use *D* or *G* as the weight factor (see above); they tend to give underestimates. This is the reason why the dimension-corrected *D* and *G* were developed (see Metsämuuronen, 2021a, 2021b). The dimension-corrected estimators behave quite optimally in the dataset: the range is narrow and the average estimates are conservative.

More specified comparisons related to DCERs. Third, the traditional estimators seem to be very instable with short tests with easy items, a narrow scale in the score, and small sample size (Figures 5–11). Also, the traditional estimators tend to deviate more from the population estimates the more extreme the difficulty is (Figure 5). This is specifically true with binary items, and the phenomenon is known from the previous simulations as well as from Case 1 and Case 2. The larger simulation dataset did not include extremely difficult or easy tests and, hence the deflation seems moderate in the dataset. Notably, in the figures to come, with polytomous items, D_2 and G_2 are used instead *D* and *G* because of their better behavior with polytomous items.

The smaller dataset with extreme item difficulties confirms the fact that the estimates of reliability by coefficient alpha are greatly affected by the item difficulty (Figure 6). As expected, coefficient rho is instable with the smallest sample size (n=25) because,

Figure 4. Range, mean, minimum, and maximum value of selected estimators of reliability with binary and polytomous items (n = 1,440 tests)



with small sample sizes, the maximum likelihood estimation produces easily extremely high values of factor loadings leading to suspiciously high estimates (see Figure 7). Of the 140 tests by $n = 25$, 6% did not get estimate at all, two were out of range (> 1.00) caused by the ultra-Haywood cases in the factor loadings ($\lambda_i = 1$), and six was suspiciously high because of the Haywood cases ($\lambda_i \approx 1$). Suspiciously high cases were not found with tests with $n \geq 50$.

In general, the estimates by *AlphaD* are more stable across the various levels of test difficulty in comparison with the traditional alpha and maximal reliability. It seems though that the estimates by the traditional estimators and DCERs differ the greatest with extremely easy items rather than with exceedingly difficult items; notably, the dataset did not include *extremely* difficult items. If we take *AlphaD* as the benchmark, with the easiest tests (average item difficulty $p > 0.80$), the estimates by the traditional alpha were deflated on average 0.27 units of reliability (32%) ranging from 0.14 units (15%) to 0.57 units (73%). Correspondingly, with the easiest tests, the estimates by maximal reliability are deflated on average by 0.15 units of reliability ranging from −0.10 units

Figure 5. Behaviour of the estimators by the test difficulty

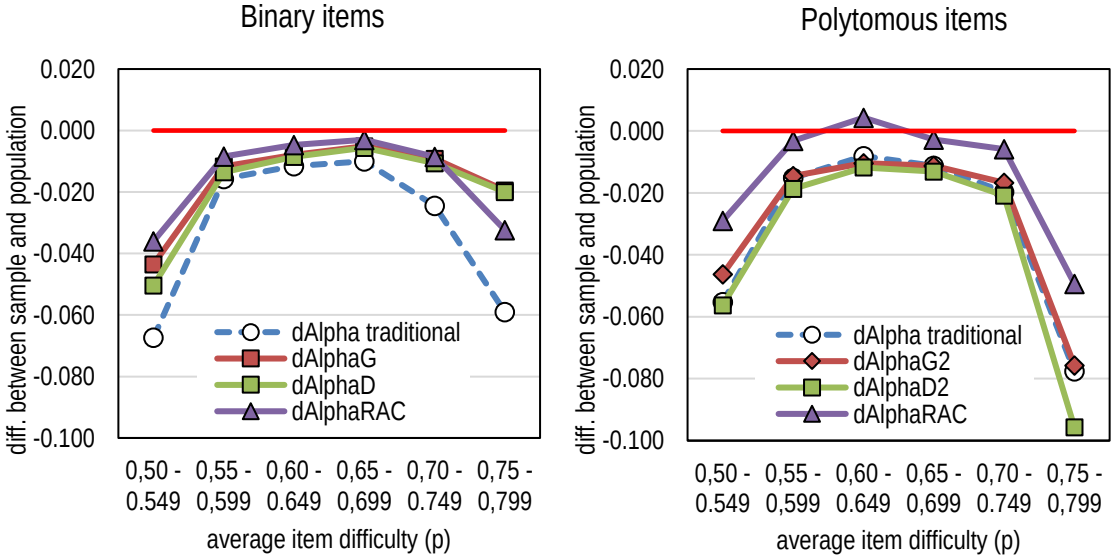


Figure 6. Behaviour of the estimators by the test difficulty in shorter tests with binary items with extreme difficulty (n = 560 estimates)

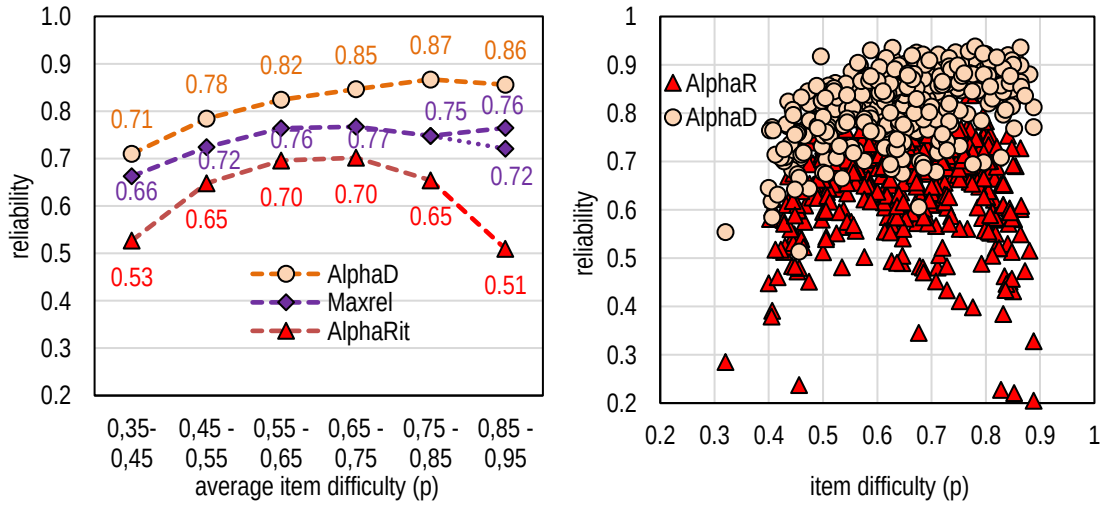
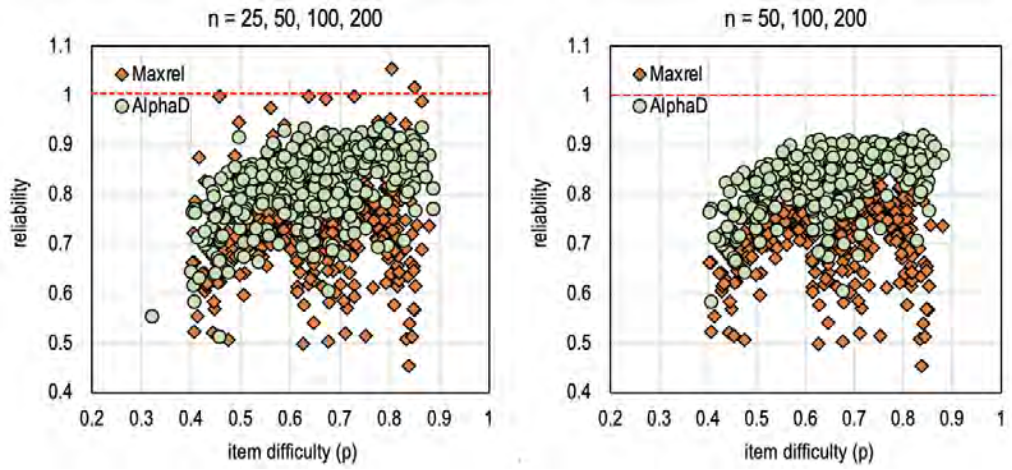


Figure 7. Behaviour of maximal reliability by the test difficulty in shorter tests with binary items with extreme difficulty (n = 560 and 420 estimates)



(the estimates by rho were higher than those by *AlphaD*) to 0.37 units (45%). That the estimates by DCERs are more stable at both ends of the scale of item and test difficulty, has a strict consequence also to the standard error of measurement (*S.E.m*). This is discussed later.

Not only are the estimates by DCER more stable at the extremes of item and test difficulty, they are more stable also when it comes to the scale of the score. In general, the estimates by reliability tend to be stable if the score includes more than 15 categories (Figure 8). Below 15 categories, the traditional estimators underestimate reliability more than DCERs. This is specifically true with binary items. The phenomenon is more obvious when using omega as the base of DCERs (see Metsämuuronen, 2022d).

The dataset including items with extreme difficulty indicate that the estimates by DCERs tend to be stable even with very short tests (Figure 9). In these cases,

Figure 8. Behaviour of the estimators by the width of the score

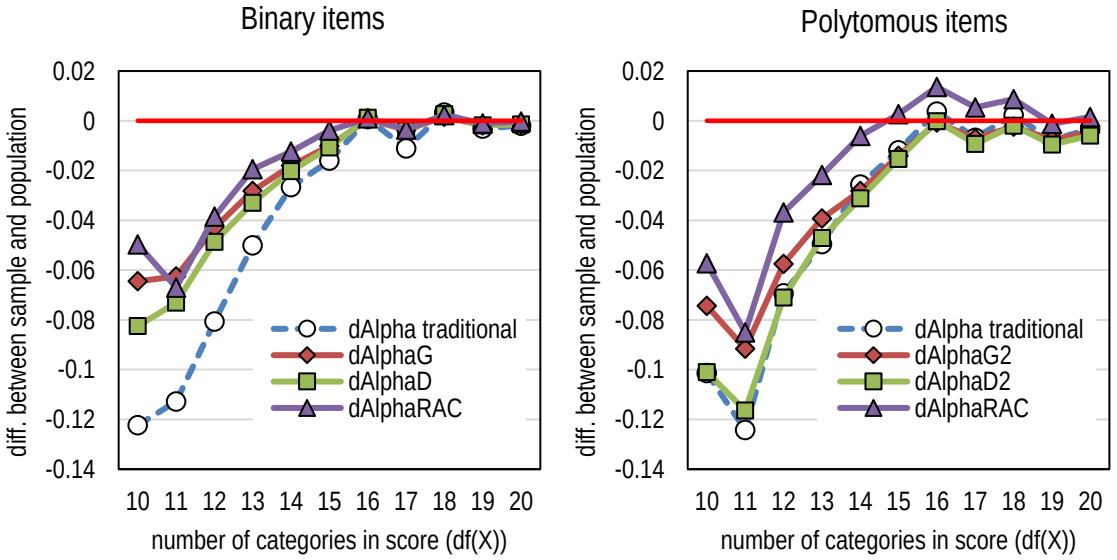
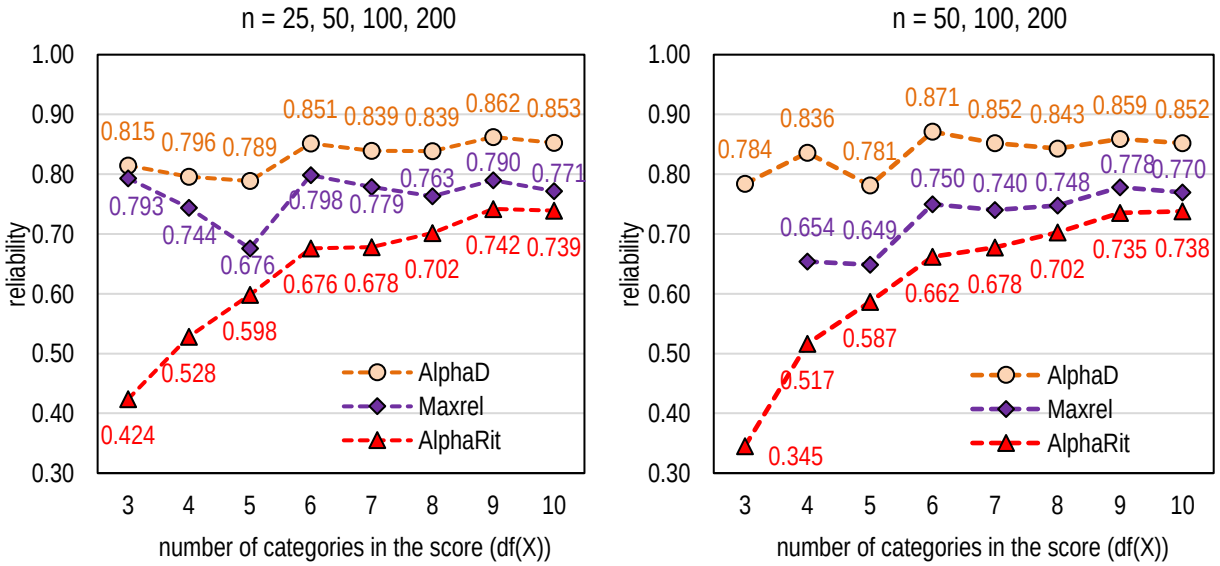


Figure 9. Behaviour of the estimators by the width of the score with very short tests with extreme difficulty levels in binary items (n = 560 and 420 estimates)



the suspiciously high estimates by maximal reliability with very small sample sizes appear to rise the average with extremely short tests.

When it comes to sample size, only four different options were in use in the analysis. Notably, the smallest, n = 25, seems to differ from the others; the traditional estimators (except rho) give notable underestimation in comparison with DCERs, specifically, with binary settings (Figure 10). When the sample size reaches or exceeds n = 100, there are no practical differences between the estimators.

Obviously, the test difficulty and width of the score are relevant factors to consider also with larger sample sizes as seen in Case 2.

With datasets with short tests and extreme difficulty levels, the factual estimates by coefficient alpha and *AlphaD* are stable over the different sample sizes (Figure 11): the average magnitude of the estimates by the traditional coefficient alpha ranges by 0.660–0.681 and by *AlphaD* by 0.824–0.842. The estimates by rho tend to get steadier by the sample size; they range 0.820–0.701 depending on the sample size.

Figure 10. Behaviour of the estimators by the sample size

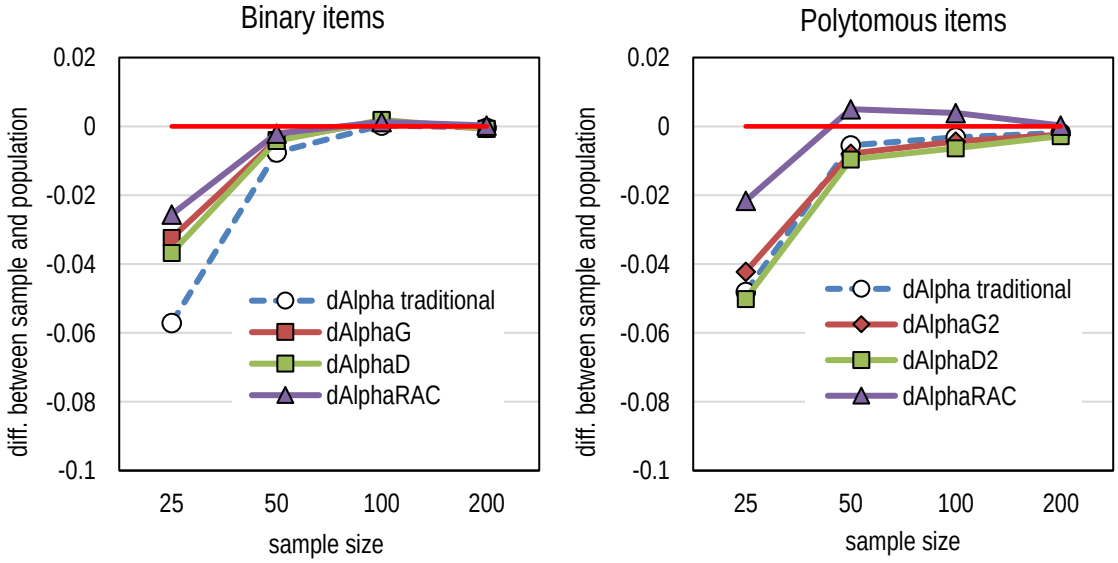
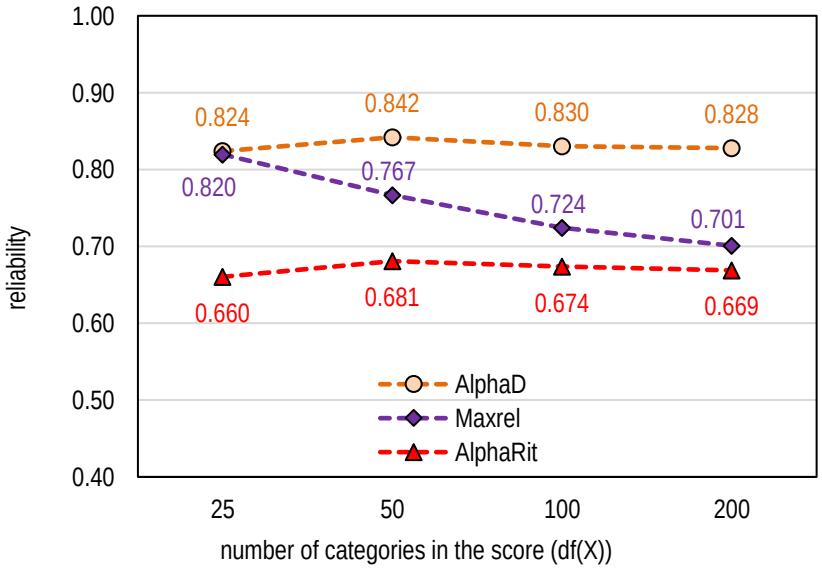


Figure 11. Average estimates of reliability by the sample size with tests of extreme difficulty levels



More detailed analysis of different estimators can be found in Metsämuuronen (2022a) where eight sources of deflation in the traditional estimators are discussed and in Metsämuuronen (2022d) where the characteristics of DCERs are typologized. In these, coefficient omega is used as the main benchmark.

Standard errors. As with Cases 1 and 2, here also the analysis is expanded to standard errors. The analysis is restricted to estimators based on coefficient alpha (*AlphaRit* and *AlphaD*). The datasets include the score variance (σ_X^2), and *S.E.m.* related to the traditional coefficient alpha is estimated by

$S.E.m.(\alpha) = \sigma_{E\_Rit} = \sigma_X\sqrt{1 - alphaRit}$ and the corresponding estimate by *alphaD* by

$S.E.m.(\alpha_D) = \sigma_{E\_D} = \sigma_X\sqrt{1 - alphaD}$.

The specific interest is in the effect of item difficulty in the estimates of *S.E.m.* From this perspective, the estimates of *S.E.m.* from both the larger dataset with non-extreme test difficulties (n = 1,440 tests) and the specific smaller dataset with short tests with extreme item difficulties (n = 560 tests) are collected in Table 4 and illustrated in Figures 12 and 13. In Table 4, the inflation rate is computed simply as the difference between the estimates based on *AlphaRit* and *AlphaD*.

Three lifts are made from Table 4 and Figures 12 and 13. First, on average, the estimates of *S.E.m.* are inflated 32%−39% with binary items by using the traditional alpha and if using *AlphaD* as the benchmark. The inflation rate in the real-life datasets seems to be notably lower than in the cases 1 and 2 above. Notably though, *AlphaD* gives us conservative estimates of reliability; the inflation would be estimated higher if, for example, Goodman–Kruskal *G* was used as the linking factor. In fact, in the larger dataset, the inflation was 43% if *AlphaG* was used as the benchmark. Second, with polytomous items, the inflation is nominal (9%), when $AlphaD_2$ is used as the benchmark. The technical reason is that even if the dimension-corrected *D* (D_2) behaves better with polytomous items than *D* does, they both give very conservative estimates with polytomous items with large scale. In these settings, the estimates by *Rit* tend to be closer the trues correlation than *D*. However, even if dimension-corrected *G* is used as the linking factor, the average inflation is 22% and with R_{PC}, 12%, that is, the inflation is notably smaller than when binary items were used. The reason is the better behavior of *Rit* with polytomous items in comparison with binary items.

Table 4. Average estimates of S.E.m. by the test difficulty

	short tests, extreme difficulty				longer tests, medium difficulty							
average item difficulty (p)	binary items				binary items				polytomous items			
	alphaRit	alphaD	inflation	n	alphaRit	alphaD	inflation	n	alphaRit	alphaD2	inflation	n
0.35−0.40	0.986	0.785	0.201	2								
0.40−0.45	1.008	0.787	0.221	23								
0.45−0.50	1.126	0.877	0.249	31								
0.50−0.55	1.279	1.001	0.278	52	2.007	1.597	0.410	10	1.945	1.831	0.114	28
0.55−0.60	1.292	0.987	0.305	59	1.981	1.544	0.437	44	1.975	1.841	0.134	138
0.60−0.65	1.162	0.875	0.287	77	2.024	1.541	0.483	84	2.029	1.872	0.157	258
0.65−0.70	1.162	0.851	0.311	97	2.097	1.530	0.566	119	2.112	1.911	0.201	450
0.70−0.75	1.129	0.783	0.346	86	2.005	1.423	0.582	60	1.963	1.765	0.198	232
0.75−0.80	1.067	0.695	0.373	62	1.894	1.290	0.604	3	1.991	1.916	0.075	14
0.80−0.85	0.898	0.532	0.365	58								
0.85−0.90	0.723	0.393	0.330	13								
Total	1.124	0.810	0.315	560	2.040	1.515	0.525	320	2.039	1.861	0.178	1120

Figure 12. Average estimates of S.E.m. by the test difficulty

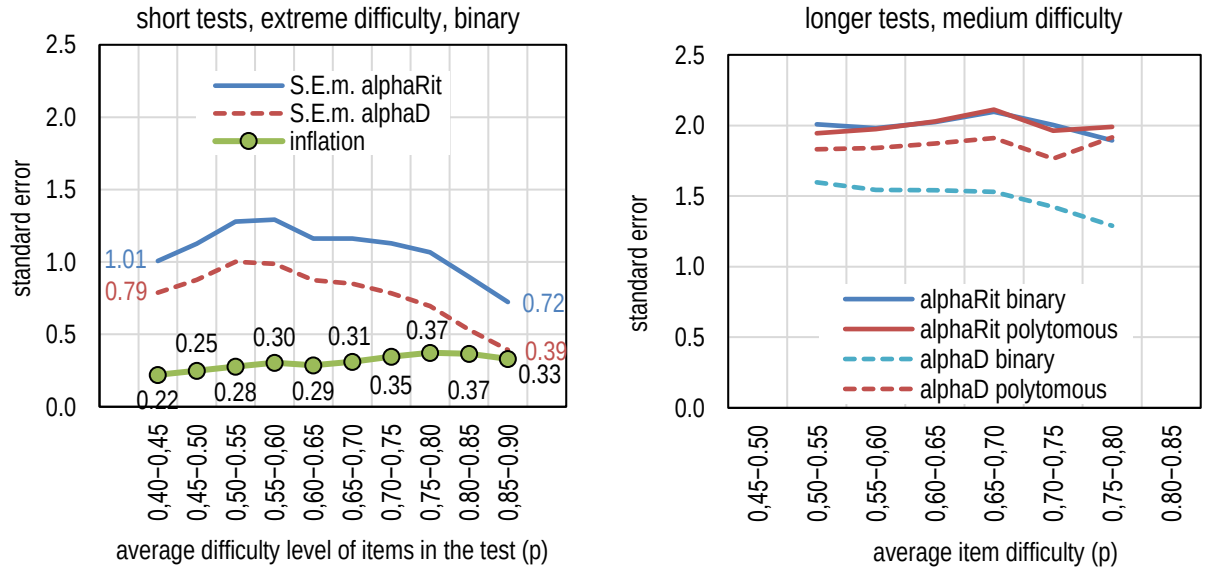
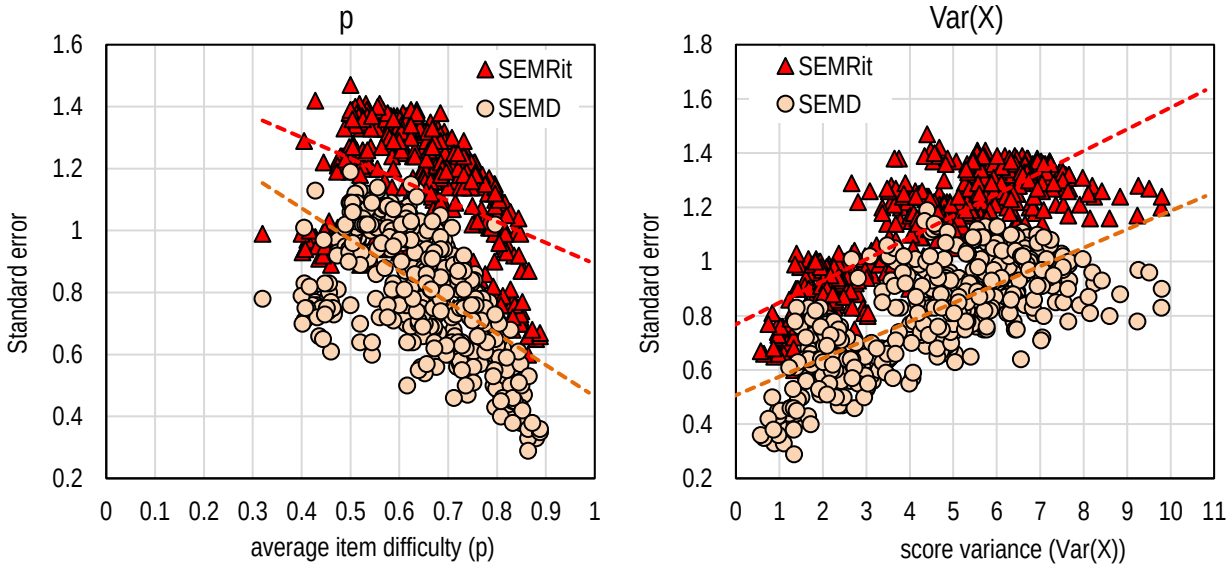


Figure 13. Average estimates of S.E.m. by the test difficulty and score variance (n = 560) with binary items



Third, with short tests of extreme difficulty levels, the inflation seems to get mildly higher by the easiness of the items, but it seems to be more stable by the score variance. Systematic studies also including extremely difficult items in this regard would be beneficial.

Conclusions and restrictions

The starting point of the study was a question why the traditional estimators of reliability such as coefficients alpha, theta, omega, and rho would not be the best option in the datasets related to achievement testing. The reason is that the traditional estimators are prone to technical or mechanical error in the estimation in such types of tests where items with extreme item difficulty are used. These kinds of items are usually administered within achievement testing. Except for some tests based on a certain standard level (such as in language testing of a certain standard level), the achievement tests usually include items with different difficulty levels—even extremely easy and difficult items may be used to cover widely the different ability levels in the target population. The items with extreme difficulty levels are those being the main reason for the deflation; the traditional estimators may be radically deflated up to 70% or 0.60 units of reliability or even more. Hence, the test construction

related to the achievement testing usually leads us to datasets that are not favoring the traditional estimators of reliability.

The root reason for the deflation in the estimates of reliability is the product–moment correlation embedded in the most used estimators of reliability. PMC is prone to give deflated estimates for the item–score association because PMC always gives underestimates when the scales of two variables are not equal as always is the case with a single item and the test score. A reasonable short-cut to reduce the negative bias in the estimates of reliability is to change the deflation-prone estimator PMC by a better behaving estimator. Previous studies have shown that certain coefficients of correlation would be reasonable alternatives for PMC. Such estimators as polychoric correlation coefficient, r-bireg and r-polyreg correlation, Somers' delta, Goodman–Kruskal gamma, dimension-corrected *D* and *G*, and attenuation-corrected *Rit* and *eta* are, if not totally deflation-free, nearly deflation-free estimators by several sources of deflation. Using these estimators instead of PMC leads us to short-cuts to the deflation-corrected estimators of reliability.

Empirical examples in this article as well as previous studies provide us with convincing facts of the behavior of the traditional estimators of reliability as well as the alternative estimators, DCERs, within achievement testing. Traditional estimators of alpha, theta, omega, and rho give *always* deflated outcomes and the deflation may be remarkable, specifically, if the scale of in the items is narrow. The empirical datasets *strongly suggest using DCERs in assessing the general quality of assessment tests instead of the traditional estimators*, if not exclusively, at least as a source of side information of whether the possibly low reliability would be caused by deflation or not. The estimates by DCERs are credibly, and for good reason, higher in magnitude than those by the traditional estimators but not overestimates; DCERs tend to reflect the population reliability more accurately than the traditional estimators.

A relevant question raised by an anonymous reviewer relates with the score variables. As the reliability is related to the variance of the score variable and, from the technical viewpoint, the benefit related to DCERs is related to the increased magnitude of the estimated score variance which is underestimated when using the traditional item–score correlation (cl. Eqs. 1, 2, and 13 and the related discussion), how this relates with the interpretation of the score variable itself (see the critical discussion in Chalmers, 2017 related to unobservable variables)? The question is whether the reliability and particularly *S.E.m.* estimated by using DCER better represent the observed (raw) score variation expected in practice under the typical reliability assumptions—or does the observed score inherent the measurement error due to the items artificially categorizing the underlying latent variable? It seems that both mechanisms are correct. The issue is discussed by Metsämuuronen (2022i, 2022j). On the one hand, the estimated population variance of the test based on the observed dataset

$$(\sigma_X^2 = \left(\sum_{i=1}^{k} \sigma_i \times \rho_{iX} \right)^2)$$

is radically deflated. Numerical examples show that, if using the estimates by deflation-corrected estimator of correlation such as $w_i = R_{PG}$, *G*, *D,* or R_{AC} instead of the traditional $w_i = Rit$, the "real" population variance may be 1.5–1.7 or even up to 2.6 times higher than what is obtained in the dataset (see Metsämuuronen, 2022i). This may be explained by another related result that the observations in the item-wise dataset which break the deterministic Guttman type of pattern (guessing in the low part and sleepiness at the high part of the scale) cause that the item difficulties are always less extreme than in the Guttman pattern (see Metsämuuronen, 2022k). If we would change the observed dataset (with stochastic error) to a form of a Guttman-patterned dataset with the same percentage of correct answer, the score variance would be remarkably higher than what was observed. This phenomenon may be worth studying more. On the other hand, from the general measurement model and error variance viewpoints (see Eq. 12);

$$\psi^2 = VAR\left(\sum_{i=1}^{k} e_i \right) = \sum_{i=1}^{k} \left(1 - w_i^2 \right)$$

the inflation in the measurement error related to a compilation of the items, i.e., the score variable, gets higher the more we have items (Metsämuuronen, 2022j). This far, the phenomenon is understandable and simple. A less simple question is how this phenomenon should be utilized in, e.g., interpreting *S.E.m.*, rectifying the attenuation, and in the meta-analytic processes (see, e.g., Sackett & Yang, 2000; Schmidt & Hunter, 2015). Studies in this respect may be valuable.

Unfortunately, the original dataset used in the empirical section did not include extremely difficult items and, hence, the behavior of the traditional estimators and DCERs remains unstudied by using this dataset. However, because of the symmetricity of coefficients of correlations in simulations (see Metsämuuronen, 2021a, 2022f) we would expect to see symmetric behavior also in the difficult extreme of item and test difficulty levels. Another question is whether the test takers behave differently with easy than difficult items. It is possible that guessing behavior with extremely difficult items changes the behavior of the estimators too. Studies in this regard may be beneficial.

The paradigm of deflation-corrected estimators of reliability is still in its infancy. Although the main lines and results are already published, systematical simulations are needed to confirm certain specificities such as their behavior with very short test, test with extreme difficulty levels, and behavior of possible alternative bases and weight factors for the estimators. Notably, the simulation dataset in Case 3 is based on one, real-world dataset; systematic simulations would enrich the discussion notably. From the viewpoint of further studies, it is largely unknown how the deflation is seen in estimators of reliability within the paradigm of generalizability theory, IRT-modelling, and nonparametric IRT modelling, and confirmatory factor analysis; estimators of reliability within these settings may include elements of deflation, but we do not know what the mechanism for deflation would be. Finally, accessible R-codes for the DECRs as well as for the new estimators of correlation (D_2, G_2, R_{AC}, E_{AC}) should be developed. However, those can be calculated manually the same manner they were calculated for this article by using common spreadsheet software, but for the applied researchers and basic user of software packages, some kind simple statistical package could be valuable to develop.

References

Aquirre-Urreta, M., Rönkkö, M., & McIntosh, C. N. (2019). A Cautionary note on the finite sample behavior of maximal reliability. *Psychological Methods. 24*(2), 236–252. https://doi.org/10.1037/met0000176

Armor, D. (1973). Theta reliability and factor scaling. *Sociological Methodology, 5*, 17–50. https://doi.org/10.2307/270831

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

Bravais, A. (1844). *Analyse Mathématique. Sur les probabilités des erreurs de situation d'un point. (Mathematical analysis. Of the probabilities of the point errors). Mémoires présentés par divers savants à l'Académie Royale des Sciences de l'Institut de France* (Memoirs presented by various scholars to the Royal Academy of Sciences of the Institute of France), *9*, 255–332. Available at https://books.google.fi/books?id=7g_hAQAACAAJ&redir_esc=y (Accessed June 28, 2023)

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322. http://dx.doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cattell, J. K. (1886). *Psychometrische Untersuchungen* (Psychometric Measurement). Engelman. Available at https://psychologie.lw.uni-leipzig.de/wundt/opera/cattell/psymtrik/PSYMETUI.htm (Accessed June 28, 2023).

Cattell, J. K. (1893). Mental Measurement. *The Philosophical Review 2*(3), 316–332. https://doi.org/10.2307/2175386

Chalmers, R. P. (2017). On misconceptions and the limited usefulness of ordinal alpha. *Educational and Psychological Measurement, 78*(6), 1056–1071. https://doi.org/10.1177/0013164417727036

Cho, E., & Chun, S. (2018). Fixing a broken clock: A historical review of the originators of reliability coefficients including Cronbach's alpha. *Survey Research, 19*(2), 23–54. Available at https://www.researchgate.net/publication/325426340_Fixing_a_broken_clock_A_historical_review_of_the_originators_of_reliability_coefficients_including_Cronbach%27s_alpha (Accessed June 28, 2023).

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–

230. https://doi.org/10.1177/1094428114555994

Cleff, T. (2019). Applied statistics and multivariate data analysis for business and economics. A modern approach using SPSS, Stata, and Excel. Springer.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*(2), 300–315. https://doi.org/10.1037/a0033805

Cramer, D. & Howitt, D. (2004). The Sage Dictionary of Statistics. A practical resource for students. SAGE Publications, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3) 297–334. https://doi.org/10.1007/BF02310555

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. http://dx.doi.org/10.1111/bjop.12046

Falk, C. F., & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment 93*(5), 445–53. https://doi.org/10.1080/00223891.2011.594129.

FINEEC (2018). *National assessment of learning outcomes in mathematics at grade 9 in 2002* (Unpublished dataset opened for the re-analysis 18.2.2018). Finnish National Education Evaluation Centre (FINEEC).

Foy, P. & LaRoche, S. (2019). Estimating standard errors in the TIMSS 2019 results. Ch. 14 in M. O. Martin, M. von Davier, & I.V.S. Mullis, (Eds.) (2019). TIMSS 2019 Technical report. Available at https://timssandpirls.bc.edu/timss2019/methods/chapter-14.html (Accessed June 28, 2023).

Gadermann A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation, PARE, 17*(3), 1–13. https://doi.org/10.7275/n560-j767

Galton, F. (1869). *Hereditary Genius*. MacMillan and Co.

Galton F (1889). Kinship and correlation. *Statistical Science*, *4*(2), 80–86. (Also, 1890 in *North American Review*, *150*, 419–431). http://doi.org/10.1214/ss/1177012581

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*(268), 732–764. https://doi.org/10.1080/01621459.1954.10501231

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications*. Allyn & Bacon.

Gulliksen, H. (1950). *Theory of mental tests*. Lawrence Erlbaum Associates Publishers.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural Equation Modeling: Present and Future — A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Scientific Software International, Inc.

Heise, D., & Bohrnstedt, G. (1970). Validity, invalidity, and reliability. *Sociological Methodology, 2*, 104–129. https://doi.org/10.2307/270785

Hoekstra, R., Vugteveen, J., Warrens, M. J., & Kruyen, P. M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology*, *22*(4), 351–364. https://doi.org/10.1080/13645579.2018.1547523

Jackson, R. W. B., & Ferguson, G. A. (1941). *Studies on the reliability of tests*. Department of Educational Research, University of Toronto.

Kane, M. T. (1986). The role of reliability in criterion-referenced tests. *Journal of Educational Measurement, 23*(3), 221–224. http://dx.doi.org/10.1111/j.1745-3984.1986.tb00247.x

Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika, 30*, 1−14. https://doi.org/10.1007/BF02289743

Kendall MG (1948) *Rank correlation methods*. First Edition. Charles Griffin & Co Ltd.

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. http://dx.doi.org/10.1007/BF02288391

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika*, *62*(2), 245–249. http://dx.doi.org/10.1007/BF02295278

Lord, F. M. (1958). Some relations between Guttman's principal component scale analysis and other psychometric theory. *Psychometrika*, *23*(4), 291–296. http://dx.doi.org/10.1002/j.2333-8504.1957.tb00073.x

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison–Wesley Publishing Company.

McDonald, R. P. (1970). Theoretical canonical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*, 1–21. http://dx.doi.org/10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Lawrence Erlbaum Associates.

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. http://dx.doi.org/10.1037/met0000144

Metsämuuronen, J. (2017). *Essentials of Research Methods in Human Sciences*. SAGE Publications.

Metsämuuronen, J. (2018). Common Framework for Mathematics—Discussions of Possibilities to Develop a Set of General Standards for Assessing Proficiency in Mathematics. *International Electronic Journal of Mathematics Education*, *13*(2), 13–39. https://doi.org/10.12973/iejme/2693

Metsämuuronen J. (2021a). Goodman–Kruskal gamma and dimension-corrected gamma in educational measurement settings. *International Journal of Educational Methodology, 7*(1), 95–118. https://doi.org/10.12973/ijem.7.1.95

Metsämuuronen, J. (2021b). Directional nature of Goodman-Kruskal gamma and some consequences. Identity of Goodman-Kruskal gamma and Somers delta, and their connection to Jonckheere-Terpstra test statistic. *Behaviormetrika, 48*/2. http://dx.doi.org/10.1007/s41237-021-00138-8

Metsämuuronen, J. (2022a). How to obtain the most error-free estimate of reliability? Eight sources of underestimation of reliability. *Practical Assessment, Research, and Evaluation, PARE*, *27*(1), Art. 10. https://doi.org/10.7275/7nkb-j673

Metsämuuronen, J. (2022b). Reliability for a score compiled from multiple booklets with equated scores. Preprint at http://dx.doi.org/10.13140/RG.2.2.20880.69120/1

Metsämuuronen, J. (2022c). Deflation-corrected estimators of reliability. *Frontiers in Psychology, 12*:748672, https://doi.org/10.3389/fpsyg.2021.748672

Metsämuuronen, J. (2022d). Typology of Deflation-Corrected Estimators of Reliability. *Frontiers in Psychology, 13*:891959. https://doi.org/10.3389/fpsyg.2022.891959

Metsämuuronen, J. (2022e). Attenuation-corrected reliability and some other MEC-corrected estimators of reliability. *Applied Psychological Measurement, 46*(8). https://doi.org/10.1177/01466216221108131

Metsämuuronen, J. (2022f). Effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability. Seeking the best options of correlation for deflation-corrected reliability. *Behaviormetrika, 49*(1), 91–130 https://doi.org/10.1007/s41237-022-00158-y

Metsämuuronen, J. (2022g). Artificial systematic attenuation in eta squared and some related consequences. Attenuation-corrected eta and eta squared, negative values of eta, and their relation to Pearson correlation. *Behaviormetrika,* https://doi.org/10.1007/s41237-022-00162-2

Metsämuuronen, J. (2022h). Directional nature of the product–moment correlation coefficient and some consequences. *Frontiers in Psychology, 13*:988660. https://doi.org/10.3389/fpsyg.2022.988660

Metsämuuronen, J. (2022i). Note on the deflation in population variance in the measurement modelling settings. Preprint.

http://dx.doi.org/10.13140/RG.2.2.31887.87202 (Accessed June 28, 2023)

Metsämuuronen, J. (2022j). Note on the inflation in error variance and standard error in the measurement modelling settings. Preprint. http://dx.doi.org/10.13140/RG.2.2.10029.15846 (Accessed June 28, 2023)

Metsämuuronen, J. (2022k). Seeking the real item difficulty: bias-corrected item difficulty and some consequences in Rasch and IRT modeling. *Behaviormetrika*. https://doi.org/10.1007/s41237-022-00169-9

Metsämuuronen, J. (2023a). How to make sense to reliability? Common language estimators of reliability and the relation of reliability to effect size. Preprint. http://dx.doi.org/10.13140/RG.2.2.31639.04000 (Accessed June 28, 2023).

Metsämuuronen, J. (2023b). Somers' delta as a basis for nonparametric effect sizes: Grissom-Kim PS, Cliff's d, and Vargha-Delaney A as specific cases of Somers delta. Preprint. http://dx.doi.org/10.13140/RG.2.2.36002.09925 (Accessed June 28, 2023).

Metsämuuronen, J. & Ukkola, A. (2019). *Alkumittauksen menetelmällisiä ratkaisuja* (*Methodological solutions of zero level assessment*). Publications 18:2019. Finnish Education Evaluation Centre. [in Finnish, abstract in English] https://karvi.fi/app/uploads/2019/08/KARVI_1819.pdf

Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London, 187*, 253–318. https://doi.org/10.1098/rsta.1896.0007

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society A. Mathematical, Physical and Engineering Sciences, 195*(262–273), 1–47. https://doi.org/10.1098/rsta.1900.0022

Pearson, K. (1903). I. Mathematical contributions to the theory of evolution. —XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society A. Mathematical, Physical and Engineering Sciences, 200*(321–330), 1–66. https://doi.org/10.1098/rsta.1903.0001

Pearson, K. (1905). On the general theory of skew correlation and non-linear regression. London. Dulau & Co. https://archive.org/details/ongeneraltheory00peargoog/page/n3

Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika*, 9(1–2), 116–139. https://doi.org/10.1093/biomet/9.1-2.116

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1–9. http://dx.doi.org/10.1111/j.1745-3984.1969.tb00654.x.

Raykov, T. (1997b). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173–184. https://doi.org/10.1177/01466216970212006

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modeling approach. *British Journal of Mathematical and Statistical Psychology, 57*(1), 21–27. http://doi.org/10.1348/000711004849295

Raykov, T., & Marcoulides, G. A. (2017). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*(1), 200–210. http://dx.doi.org/10.1177/0013164417725127

Raykov, T., West, B. T., & Traynor, A. (2015). Evaluation of coefficient alpha for multiple component measuring instruments in complex sample designs. *Structural Equation Modeling, 22*(3), 429-438. http://dx.doi.org/10.1080/10705511.2014.936081.

Revelle, W., & Condon, D. M. (2018). *Reliability from a to ω: A tutorial.* http://doi.org/10.31234/osf.io/2y3w9

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Education Review, 9,* 99–103.

Sackett, P. R., Lievens, F., Berry, C., M. & Landers, R. N. (2007). A cautionary note on the effect of range restriction on predictor intercorrelations. *Journal of Applied Psychology, 92*(2) 538–544. http://doi.org/10.1037/0021-9010.92.2.538

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). SAGE Publications. https://dx.doi.org/10.4135/9781483398105

Schult, J. & Sparfeldt, J.R. (2016). Reliability and validity of PIRLS and TIMSS. *European Journal of Psychological Assessment, 34*(4), 258–269. https://doi.org/10.1027/1015-5759/a000338

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

Silver, N. C. (2008). Attenuation. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Methods*. Sage Publications, Inc. https://dx.doi.org/10.4135/9781412963947.n24

Somers, R. H. (1962). A new asymmetric measure of correlation for ordinal variables. *American Sociological Review, 27*(6), 799–811. http://dx.doi.org/10.2307/2090408

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101. http://dx.doi.org/10.2307/1422689

Spearman, C. (1910). Correlation computed with faulty data. *British Journal of Psychology, 3*(3), 271–295. http://dx.doi.org/10.1111/j.2044-8295.1910.tb00206.x

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. http://dx.doi.org/10.3389/fpsyg.2016.00769

Wundt, M. (1862). *Beiträge zur Theorie der Sinneswahrnehmung* (Effect on Theory of observing senses). C.F. Winter'sche Verlagshandlung.

Yang, Y. & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st Century? *Journal of Psychoeducational Assessment*, *29*(4), 377–392. http://dx.doi.org/10.1177/0734282911406668

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1), 21–29. http://dx.doi.org/10.22237/jmasm/1177992180

Citation:
Metsämuuronen, J. (2023). Seeking the real reliability: Why the traditional estimators of reliability usually fail in achievement testing and why the deflation-corrected coefficients could be better options. *Practical Assessment, Research, & Evaluation*, 28(10). Available online: https://scholarworks.umass.edu/pare/vol28/iss1/10/

Corresponding Author:
Jari Metsämuuronen
University of Turku, Turku Research Institute for Learning Analytics
Finnish National Education Evaluation Centre (FINEEC)
Email: jari.metsamuuronen [at] gmail.com

Appendix 1. Estimation of the coefficients of correlation used in DCERs

Rit

In IBM SPSS, the syntax for *R*it is CORRELATIONS /VARIABLES=g X or CROSSTABS /TABLES=item BY Score /STATISTICS=CORR. In SAS, the command PROC CORR provides Rit. Correspondingly, in R, *Rit* is calculated by *cor(datg, datX)* (see, e.g., https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/#between-two-variables). For the empirical section, *Rit* was calculated by using basic spreadsheet software (MS-EXCEL).

D

In IBM SPSS, the syntax for *D* is CROSSTABS /TABLES=item BY Score /STATISTICS=D. The option "Score dependent" is selected. In SAS, the command PROC FREQ provides D by specifying the TEST statement by DELTA, SMDCR options. Correspondingly, in RStudio, *D* is calculated by *SomersDelta(x, y = NULL, direction = c(*"row", "column"*), conf.level = NA, ...)* (see https://rdrr.io/cran/DescTools/man/). For the empirical section, *D* was calculated by using IBM SPSS software.

G

In traditional software packages such as IBM SPSS, for instance, the syntax for *G* is CROSSTABS /TABLES=item BY Score /STATISTICS=GAMMA. In SAS, the command PROC FREQ provides *G* by specifying the TEST statement by GAMMA, SMDCR options. Correspondingly, in RStudio *G* is calculated by *GoodmanKruskalGamma(x, y = NULL, conf.level = NA, ...)* (see https://rdrr.io/cran/DescTools/man/). For the empirical section, *G* was calculated by using IBM SPSS software.

Attenuation-corrected PMC (RAC)

RAC is the proportion of the observed item–score correlation (ρ_{gX}^{Obs}) of maximal correlation (ρ_{gX}^{Max}) possible to obtain with the observed *g* and *X:* $\rho_{AC} = \frac{\rho_{gX}^{Obs}}{\rho_{gX}^{Max}} = \frac{\rho_{gX}}{\rho_{gX}^{Max}}$. The maximum values of *Rit* in the given dataset are obtained when the correlation is calculated between variables *g* and *X* after they are *ordered independently*. In the traditional software packages such as IBM SPSS, for instance, the syntax for *Rit* is CROSSTABS /TABLES=item BY Score /STATISTICS=CORR. In SAS, the command PROC CORR provides *Rit*. In R, *Rit* is calculated by *cor(datg, datX)* (see, e.g., https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/#between-two-variables). In R, the variables (vectors) can be sorted by a command *sort (x)* #. For the empirical section, both the maximal and observed *Rit* were calculated manually by using a spreadsheet software (MS-EXCEL).

Principal component and factor loadings

In IBM SPSS, the principal component loadings are calculated by the command FACTOR /VARIABLES g1 g2 g3 g4 g5 /MISSING LISTWISE /ANALYSIS g1 g2 g3 g4 g5 /PRINT INITIAL EXTRACTION /CRITERIA FACTORS(1) ITERATE(25) /EXTRACTION PC /ROTATION NOROTATE /METHOD=CORRELATION. Parallel, the factor loading with maximum likelihood estimation is calculated by the command FACTOR

/VARIABLES g1 g2 g3 g4 g5 /MISSING LISTWISE /ANALYSIS g1 g2 g3 g4 g5 /PRINT INITIAL EXTRACTION /CRITERIA FACTORS(1) ITERATE(25) /EXTRACTION ML /ROTATION NOROTATE.
The table "factor matrix" is selected.
The procedures for principal component and factor analyses with SAS can be found at
https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_factor_examples01.htm

In R, the **princomp( )** function produces an unrotated principal component analysis.
https://www.statmethods.net/advstats/factor.html gives a syntax as follows:

```
# Pricipal Components Analysis
# entering raw data and extracting PCs
# from the correlation matrix
fit <- princomp(mydata, cor=TRUE)
summary(fit) # print variance accounted for
loadings(fit) # pc loadings
plot(fit,type="lines") # scree plot
fit$scores # the principal components
biplot(fit)
```

Parallel, syntax for the factor analysis is

```
# Maximum Likelihood Factor Analysis
# entering raw data and extracting 3 factors,
# with varimax rotation
fit <- factanal(mydata, 3, rotation="varimax")
print(fit, digits=2, cutoff=.3, sort=TRUE)
# plot factor 1 by factor 2
load <- fit$loadings[,1:2]
plot(load,type="n") # set up plot
text(load,labels=names(mydata),cex=.7) # add variable names
```