Applied Psychological Measurement
2022, Vol. 0(0) 1–18
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216221108131
journals.sagepub.com/home/apm
SAGE

Attenuation-Corrected Estimators of Reliability

Jari Metsämuuronen[1,2]

Abstract

The estimates of reliability are usually attenuated and deflated because the item–score correlation (ρ_{gX}, *Rit*) embedded in the most widely used estimators is affected by several sources of mechanical error in the estimation. Empirical examples show that, in some types of datasets, the estimates by traditional alpha may be deflated by 0.40–0.60 units of reliability and those by maximal reliability by 0.40 units of reliability. This article proposes a new kind of estimator of correlation: attenuation-corrected correlation (R_{AC}): the proportion of observed correlation with the maximal possible correlation reachable by the given item and score. By replacing ρ_{gX} with R_{AC} in known formulas of estimators of reliability, we get attenuation-corrected alpha, theta, omega, and maximal reliability which all belong to a family of so-called deflation-corrected estimators of reliability.

Keywords

reliability, attenuation in reliability, attenuation in correlation, attenuation-corrected estimators of reliability, deflation-corrected estimators of reliability

Introduction—Attenuation and Deflation in Correlation and Reliability

The reliability of a test score variable has interested scholars for more than a century and for good reasons. Reliability serves four main purposes: it is used in estimating error in the score of an individual test taker, that is, in indicating the (overall) measurement error in the test score (Gulliksen, 1950), in assessing the (overall) quality of the measurement (e.g., Metsämuuronen, 2017), in correcting the attenuation of the estimates of regression or path models (e.g., Cole & Preacher, 2014), and in correcting the attenuation in correlations in validity studies and meta-analyses (e.g., Schmidt & Hunter, 2015). In all cases, we want to obtain as accurate estimate of reliability as possible.

Two terms related to correlation and reliability are worth highlighting here: attenuation and deflation. Usually, attenuation refers to underestimation as a natural consequence of random errors in the measurement, and deflation refers to underestimation caused by artificial systematic errors during the estimation (see the discussion of the terms in, e.g., Chan, 2008; Gadermann et al., 2012;

[1]Finnish Education Evaluation Centre (FINEEC), Helsinki, Finland
[2]Centre for Learning Analytics, University of Turku, Turku, Finland

Corresponding Author:
Jari Metsämuuronen, Finnish Education Evaluation Centre (FINEEC), (Hakaniemenranta 6), Helsinki 00531, Finland.
Email: jari.metsamuuronen@karvi.fi

Metsämuuronen, 2022a; Revelle & Condon, 2018). These are not always easy to separate from each other and, hence, both terms are used during the article. Although deflation may be closer the focus in this article, the term attenuation is mainly used. These concepts are connected to a new concept called here "mechanical error in estimates of correlation" (MEC), that is, a characteristic of estimators of correlation to underestimate the true correlation because of technical or mechanical reasons.

This article discusses, first, how the attenuation and deflation in correlation and in reliability are intertwined and exacerbated: attenuation and deflation in correlation are seen to be the *reasons* for the attenuation and deflation in reliability. Second, some conceptual and practical options are discussed about how to reduce the deflation in the estimates of reliability.

Attenuation, Deflation, Restriction in Range, and MEC in Correlation

Attenuation in correlation has been discussed for more than a century starting from Pearson (1903) and Spearman (1904). The phenomenon has been studied, specifically, by scholars working on restriction of range (RR; see the literature in Sackett & Yang, 2000; Sackett et al., 2007; Schmidt et al., 2008). RR is phenomenon that when only a portion of the range of values of the variable is actualized in the sample, such as when only the best performing students from the population apply to a very demanding study program causing that the variance in the entrance test is reduced remarkably, this leads to inaccuracy of the estimates used in prediction the performance (see illustrations of different patterns of RR in Sackett & Yang, 2000). Another area where the attenuation of correlation is considered in detail are validity studies and meta-analytic studies (see literature in, e.g., Schmidt & Hunter, 2015; Schmidt et al., 2008). Many options to correct attenuation have been offered, specifically, related to concurrent validity of the scores (see a typology and history in Sackett & Yang, 2000; Sackett et al., 2007; see also Schmidt et al., 2008).

Deflation or inaccuracy in the estimates of correlation has been noticed in several simulations (see, e.g., in Martin, 1973, 1978; Metsämuuronen, 2021a, 2022b; Olson, 1980). Even if there is no traditional manifestation of RR in a test, the product-moment correlation coefficient itself (PMC; Pearson, 1896) is deflated because PMC is very vulnerable to several sources of MEC (see Metsämuuronen, 2021a, 2022b). It is known that PMC cannot reach the ultimate perfect correlation unless the variables have equal number of categories (see algebraic reasons in, e.g., Metsämuuronen, 2017); this is the technical reason for attenuation. This obvious attenuation or deflation is simple to observe if we have two identical continuous variables which are truncated such that one is dichotomized (g) and the other is polytomized (X). The magnitude of the estimates of the correlation between these different manifestations of the same variable by PMC cannot reach the obvious (latent) perfect correlation but, instead, the highest value depends on several factor even without RR. Some of these sources of MEC are well-known such as the number of categories in g and X and the cut-off where the dichotomization and polytomization has been made (see simulations, e.g., in Metsämuuronen, 2020a, 2022b).

Attenuation and deflation in correlation has a strict relevance in measurement modeling settings where the dimensions of items (g) and score (X) differ from each other in an obvious manner. Specifically, Schmidt and Hunter (1999) have pointed out that we should try to embed corrections of attenuation as part of our estimations of measurement error. This is an understandable and relevant point because the obvious attenuation in PMC has a strict and remarkable effect on the estimates of reliability. This is illustrated later by empirical examples. Although the issue has been known for tens of years, and some solutions have been offered for the practical use, the conceptualization of MEC in the measurement models is somewhat undeveloped (see an attempt in Supplement Appendix 1). This article aims to illustrate the effect of attenuation and deflation in the estimators of reliability and to offer some practical solutions for the problem.

Attenuation and Deflation in Reliability

Brown (1910) and Spearman (1910) may be the first scholars to connect attenuation of the estimates of correlation with the estimates of reliability—although in an opposite way to the interest in this article. Originally, the first estimator of reliability, Brown–Spearman parallel reliability coefficient (BS; of the reasoning for the unconventional order of the inventors, see Cho & Chun, 2018), was invented to get a better approximation of correlation in the case of "faulty data" (see also Guttman, 1945). In this article, the viewpoint is flipped: the flaws in correlation coefficient are, factually, the elementary reason for the mechanical underestimation in reliability. Although flipping the viewpoint, we note Brown's and Spearman's remarkable role in initiating the discussion of measurement error and reliability as we have today.

Another important scholar in the history pointing to the mechanical error in the estimates of reliability was Louis Guttman whose seminal study of different lower bounds of reliability (Guttman, 1945) is still valid: the true population reliability is always higher than the observed reliability by the coefficient we know today as coefficient alpha (α; chronology, Guttman, 1945; Jackson & Ferguson, 1941; Kuder & Richardson, 1937) or Cronbach's alpha (Cronbach, 1951). Novick and Lewis (1967) continued Guttman's work and showed that underestimation is caused by deviation of the (essential) tau-equivalency: if all items have (essentially) identical true value (*tau*), alpha will not underestimate reliability (see the discussion also in Raykov & Marcoulides, 2017).

Since Guttman (1945) and Novick and Lewis (1967), numerous studies have handled the underestimation in reliability and, specifically, in α. Attenuation in α has been connected to a simplified assumption of the classical test theory including violations in tau–equivalence, unidimensionality, and uncorrelated errors (e.g., Green & Yang, 2009; Trizano-Hermosilla & Alvarado, 2016). Among others, Gadermann et al. (2012), Metsämuuronen (2017, 2020a, 2021a, 2022b), Zumbo et al. (2007) have discussed a different type of reason for the underestimation in reliability: technical underestimation of correlation by PMC discussed above.

Notably, PMC is embedded in most of the widely used estimators of reliability. PMC in the form of item–score correlation (ρ_{gX}) is *visible* in such classic estimators of reliability as BS, Flanagan–Rulon formula (Rulon, 1939), Kuder–Richardson formulas 20 and 21 (KR20, KR21; Kuder & Richardson, 1937), coefficient alpha, lambda family ($\lambda_1 - \lambda_6$, Guttman, 1945), and the greatest lower bound of reliability based on Guttman's λ_4 (GLB; Jackson & Agunwamba, 1977; Woodhouse & Jackson, 1977). Common to these estimators is that the variance of the test score (σ_X^2) inherited from the basic definition of reliability ($REL = \sigma_T^2/\sigma_X^2 = 1 - \sigma_E^2/\sigma_X^2\varepsilon$) is strictly expressed in the formula, and σ_X^2 can be expressed using item variances (σ_g^2) and ρ_{gX}: $\sigma_X^2 = \left(\sum_{g=1}^{k} \sigma_g \times \rho_{gX}\right)^2$ (Lord et al., 1968) where k refers to number of partitions or items in the compilation. Then, coefficient alpha, as an example, can be expressed as

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_g^2}{\left(\sum_{g=1}^{k} \sigma_g \times \rho_{gX}\right)^2}\right) \tag{1}$$

(Lord et al., 1968) where PMC is visible.

PMC is *embedded* in such estimators as Armor's theta (ρ_{TH}; Armor, 1973; see also Kaiser & Caffrey, 1965; Lord, 1958)

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

where λ_i are principal component loadings of the (first) principal component and which maximizes the estimates by coefficient alpha (Greene & Carmines, 1980), McDonald's omega total, later just omega ($\rho_{\omega_{Total}} = \rho_\omega$; Heise & Bohrnstedt, 1970; McDonald, 1970)

$$\rho_\omega = \frac{\left(\sum_{i=1}^{k}\lambda_i\right)^2}{\left(\sum_{i=1}^{k}\lambda_i\right)^2 + \sum_{g=1}^{k}\left(1-\lambda_i^2\right)} \tag{3}$$

and maximal reliability (ρ_{MAX}; see the conceptualization in Li, 1997; Li et al., 1996; and different estimators in, e.g., Aquirre-Urrita et al., 2019; Cheng et al., 2012; Raykov, 2004)

$$\rho_{MAX} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k}\left(\lambda_i^2/\left(1-\lambda_i^2\right)\right)}} \tag{4}$$

(e.g., Cheng et al., 2012). Relation of ρ_{gX} and the principal component loadings (in ρ_{TH}) and factor loadings (in ρ_ω and ρ_{MAX}) is understandable because the loadings are, essentially, correlations between an item and a score variable (e.g., Cramer & Howitt, 2004; Kim & Mueller, 1978; Yang, 2010); λ_i is essentially ρ_{gX}.

Attenuation and Deflation in Reliability in Practical Testing Settings

The consequence of the attenuation in ρ_{gX} is that, using the traditional estimators of reliability, the estimates are always negatively biased. While ρ_{gX} is severely attenuated and deflated, specifically, with items of extreme difficulty levels, we can predict that if the test is very difficult or very easy for the target group, or if the items are incrementally difficult including both easy and difficult items, the estimate of reliability would be attenuated remarkably because of MEC. If we use attenuation-corrected ρ_{gX} or if replacing ρ_{gX} in the estimators of reliability with another coefficient of correlation shown to be less MEC-affected, we get attenuation- or MEC-corrected estimators of reliability with a possibly remarkable reduction of attenuation in reliability. The phenomenon is illustrated here by two empirical examples, and the factual estimators are discussed in section Attenuation and other deflation-corrected estimators of reliability.

The first example comes from Metsämuuronen and Ukkola (2019). They used a very easy subtest of preconditions in mother language at the beginning of the first year in school as a part of a National level assessment of learning outcomes (n = 7770); 72% of the test takers got the highest possible score in 8-item, 11-point screening test for basic skills of understanding auditive instructions. The magnitude of the estimate by coefficient alpha was α = 0.25 and by maximal reliability $\rho_{MAX} = 0.48$. The outcome was checked by a MEC-corrected estimator (α_D; see later equation (12)) by replacing ρ_{gX} in the formula of alpha by Somers D (Somers, 1962), known to be more resistant to MEC than PMC, especially, with binary items (see Metsämuuronen, 2020a, 2020b, 2021a). The magnitude of the estimate appeared to be $\alpha_D =$ 0.86. Hence, in comparison with α_D, the estimate by α was deflated from 0.86 to 0.25, that is, 0.61 units of reliability (or 71%; see this interpretation in Gadermann et al., 2012) and maximal reliability was deflated 0.38 units of reliability (46%). Deflation of a same magnitude was found also in Gadermann et al.'s (2012) dataset where, by using ordinal alpha (α_{ORD}), another kind of MEC-corrected estimator based on polychoric correlation (R_{PC}; Pearson, 1900, 1913), the estimate by α was deflated from 0.85 (α_{ORD}) to 0.46 (α), that is, 0.39 units of reliability (46%). In both examples the attenuation in α (as well as in ρ_{MAX}) is remarkable and worth noting.

The obvious reason for the remarkably higher estimate by α_D and α_{ORD} than by the traditional α and ρ_{MAX} is that PMC and factor loadings are severely affected by MEC in items with extreme difficulty level while D and R_{PC} are less affected by MEC in the binary case (see Metsämuuronen, 2021a, 2022b). In the case of Metsämuuronen and Ukkola (2019), the magnitude of estimate would have been even higher if R_{PC} or Goodman–Kruskal gamma (G; Goodman & Kruskal, 1954) was used instead of D in the formula because the estimates by D are usually more conservative than those by G (see a re-analysis of the dataset in Metsämuuronen, 2022a). Some exceptions of the conservativeness of D are discussed by, for example, Metsämuuronen (2021b), and, in the case of binary items, the difference between the estimates by R_{PC} and G is nominal (Metsämuuronen, 2021a).

Research Question

Seeking the most accurate estimate of reliability is important for all four main uses of reliability discussed above. Attenuation and deflation in reliability is technical and caused, mainly, by the mechanical error in ρ_{gX}. A relevant question is, how to solve the issue of attenuation in ρ_{gX} and how this could be utilized in estimating reliability.

While there are some MEC-corrected estimators available based on changing the whole estimator of correlation (see Gadermann et al., 2012; Metsämuuronen, 2020b, 2021a, 2021b, 2022a; Zumbo et al., 2007; some are discussed below), this article studies the option of a relevant attenuation correction for ρ_{gX} as a solution. A simple correction of attenuation for ρ_{gX} is proposed first after which a family of attenuation-corrected estimators of reliability is proposed and numerical examples are given of their behavior in three forms of datasets: (1) a dataset of extreme difficulty level, (2) a dataset of incremental difficulty level, and (3) a larger simulation based on a dataset of 1440 real-world tests. The conceptual discussion related to the attenuation-corrected estimators of reliability is incorporated in Supplement Appendix 1 (see also Metsämuuronen, 2022a, 2022b).

Attenuation- and Other Deflation-Corrected Estimators of Reliability

In what follows, estimators of reliability based on MEC-corrected measurement model (see Supplement Appendix 1) are divided into two categories as discussed above. Estimators based on replacing ρ_{gX} by a totally different coefficient are called MEC-corrected estimators of reliability (MCER) and estimators based on correcting ρ_{gX} by a relevant attenuation correction, are called attenuation-corrected estimators of reliability (ACER). These together form an extended family of deflation-corrected estimators of reliability (DCER). Although the focus is on ACERs, some MCERs are introduced first as benchmarks.

Selected MEC-Corrected Estimators of Reliability as Benchmarks

Metsämuuronen (2021a, 2022a, 2022b), specifically, discuss using the formula (1) as a basis of estimating MEC-corrected alpha where the error-causing ρ_{gX} is replaced by a totally different estimator that would be less affected by MEC. Some options of this kind of coefficients have been suggested in earlier works: R_{PC} (see Gadermann et al., 2012; Metsämuuronen, 2022a; Zumbo et al., 2007), G and dimension-corrected G (G_2; Metsämuuronen, 2021a, 2022a) as well as D and dimension-corrected D (D_2; Metsämuuronen, 2020b, 2021a, 2022a; Metsämuuronen & Ukkola, 2019). Of these, in simulations, R_{PC}, G, and G_2 appear to be MEC-free in many conditions of MEC affecting obviously in PMC (Metsämuuronen, 2021a, 2022b). Several other coefficients may be found potential as substitutes for ρ_{gX} (see some options in Metsämuuronen, 2020a, 2022b, Moses, 2017).

Let us denote a general weight factor between an item g_i and the latent variable θ by w_i. If we apply the estimator based on alpha (equation (1)), select the raw score (X) as the manifestation of θ, and vary w_i, a theoretical form of MCERs based on the formula of alpha is $\rho_{\alpha\_wiX} = \alpha_{wiX} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times w_{iX}\right)^2}\right)$. Selecting $w_{iX} = \rho_{gX}$ leads us to the traditional alpha while some MCERs based on different estimators of correlation can be based on R_{PC}, G, G_2, D, and D_2

$$\rho_{\alpha\_RPCiX} = \alpha_{RPCiX} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times R_{PCiX}\right)^2}\right) \tag{5}$$

$$\rho_{\alpha\_GiX} = \alpha_{GiX} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times G_{iX}\right)^2}\right) \tag{6}$$

$$\rho_{\alpha\_G_2iX} = \alpha_{G_2iX} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times G_{2iX}\right)^2}\right) \tag{7}$$

$$\rho_{\alpha\_DiX} = \alpha_{DiX} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times D(g|X)_{iX}\right)^2}\right) \tag{8}$$

$$\rho_{\alpha\_D_2iX} = \alpha_{D_2iX} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times D_{2iX}\right)^2}\right) \tag{9}$$

where R_{PCiX}, G_{iX}, G_{2iX}, $D\ (g|X)_{iX}$, and D_{2iX} represent different types of coefficients of correlation between an item and the score variable where the magnitude of MEC is lower than in PMC. Of these, the estimator based on R_{PC} (equation (5)) refers to an unreachable and *theoretical* score (see the discussion in Chalmers, 2017). The estimators based on G and D (equations (6) and (8)) have concrete interpretations in reflecting the proportion of logically ordered test takers in the dataset (see Metsämuuronen, 2021a, 2021b). The estimators based on G_2 and D_2 (equations (7) and (9)) correct the obvious underestimation of association by G and D in the case of polytomous items (Metsämuuronen, 2020b, 2021a) and, hence, the estimates are more liberal than those based on G and D. In simulations (e.g., Metsämuuronen, 2021a, 2022b), all these have shown to be less-MEC-defected options than PMC. Later, in the numerical section, these estimators are used as benchmarks to ACERs. Similar types of estimators could also be proposed based on equations (2)–(4) although these are not discussed here (see, however, Metsämuuronen, 2022a). Notably, Zumbo et al.'s (2007) ordinal alpha and ordinal theta (Gadermann et al., 2012) are based on this idea although leading to different practicalities: for the calculations, the matrix of PMCs is replaced by a matrix of R_{PC}s.

Attenuation-Corrected Correlation as a Substitute of ρ_{gX} in the Estimators of Reliability

Specific types of DCERs are obtained if, in the estimators, PMC is replaced by an attenuation-corrected PMC. Attenuation in correlation have been studied since the dawn of estimators as discussed above. The well-known corrections based on works of Pearson (1903 with notes by Aitken, 1934 and Lawley, 1943) and Thorndike (1949) are based on correcting attenuation when restriction occurs in *one* variable, that is, in the score variable. This kind of attenuation correction is used, specifically, when selecting personnel or students based either directly or indirectly (that is, as a part of some other criteria) on their performance in a test. The idea is to enhance the concurrent validity of the test score of this restricted sample by altering it either by knowing or modeling the behavior of unrestricted population variance (see the mechanics in, e.g., Sackett & Yang, 2000; Schmidt et al., 2008). These approaches to attenuation correction do not seem usable in item analysis settings. Hence, another logic is proposed as an option for measurement modeling settings.

To propose another type of attenuation correction, we recall that the correlation between an item and a score given the dataset cannot exceed the limit specified by the observed responses in the item and the score. Namely, given the score and the observed response pattern in the item, the score variance (σ_X^2) and item variance (σ_g^2) are fixed. Recalling the basic formula of PMC $(\rho_{gX} = \sigma_{gX}/\sigma_g\sigma_X)$, the only element affecting the magnitude of correlation is the item–score covariance (σ_{gX}). The maximum value of σ_{gX} is obtained when g and X are in the *same order*. Then, a simple attenuation correction related to ρ_{gX} (ρ_{AC}, R_{AC}) is to proportion the observed correlation (ρ_{gX_Obs}) with the maximal possible correlation (ρ_{gX_Max}) given the observed score and item

$$\rho_{AC} = R_{AC} = \frac{\rho_{gX\_Obs}}{\rho_{gX\_Max}} \tag{10}$$

R_{AC} proposed here is not restricted to measurement modeling settings; g and X refer to general variables with a narrower and wider scale, respectively. Calculation of R_{AC} is illustrated later with numerical examples.

Attenuation-Corrected Estimators of Reliability

If we apply the estimators (1), (2), (3), and (4), and we use R_{AC} as the manifestation of the linking element w_i, but we do not fix the manifestation of θ, we get (a theoretical) attenuation-corrected alpha.

$$\rho_{\alpha\_wi\theta} = \rho_{\alpha\_RACi\theta} = \alpha_{RAC} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times R_{ACi\theta}\right)^2}\right) \tag{11}$$

attenuation-corrected theta

$$\rho_{TH\_wi\theta} = \rho_{TH\_RACi\theta} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} R_{ACi\theta}^2}\right) \tag{12}$$

attenuation-corrected omega

$$\rho_{\omega\_wi\theta} = \rho_{\omega\_RACi\theta} = \frac{\left(\sum_{i=1}^{k} R_{ACi\theta}\right)^2}{\left(\sum_{i=1}^{k} R_{ACi\theta}\right)^2 + \sum_{i=1}^{k}\left(1 - R_{ACi\theta}^2\right)} \tag{13}$$

and attenuation-corrected ρ_{MAX}

$$\rho_{MAX\_wi\theta} = \rho_{MAX\_RACi\theta} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k}\left(R_{ACi\theta}^2 / \left(1 - R_{ACi\theta}^2\right)\right)}} \tag{14}$$

Using the estimators (2), (3), and (4) outside of their original context is, obviously, debatable; here a stand is taken that they *could* be used as stand-alone estimators even without their original contexts related to principal component- and factor analysis. Alternatively, the estimators (12), (13), and (14) may be taken as outputs of renewed procedures in the factor- and principal component analysis where λ_i is an attenuation-corrected loading. In all cases, the magnitude of the attenuation-corrected correlation is higher than the observed loading and, consequently, the attenuation-corrected reliability is expected to be higher than the reliability obtained by using traditional estimator.

If we apply equation (10) to estimators (11) to (14), we get ACERs based on R_{AC}. In the extreme cases where all items can discriminate the test takers in a deterministic manner, that is, in the case of $R_{ACi} = R_{ACj} = 1$, the ACERs based on the forms of theta (equation (12)) and omega (equation (13)) would lead to perfect reliability irrespective of item variances: $\rho_{TH\_RACX} = \frac{k}{k-1}\left(1 - \frac{1}{k}\right) \equiv 1$ and $\rho_{\omega\_RACX} = \frac{(k)^2}{(k)^2+0} \equiv 1$. In the case, estimator (11) reaches the value $\alpha_{RAC} = 1$ only when all item variances are equal. Then, $\widehat{\alpha}_{RAC} = \frac{k}{k-1}\left(1 - \frac{k\sigma^2}{(k\sigma)^2}\right) = \frac{k}{k-1}\left(1 - \frac{1}{k}\right) \equiv 1$. Otherwise, the maximum value is $\alpha_{RAC}^{MAX} = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k}\sigma_i^2 / \left(\sum_{i=1}^{k}\sigma_i\right)^2\right)$. Notably, in the case of deterministic discrimination (in any of the items), a coefficient based on ρ_{MAX} (equation (14)) could not be used due to mathematical reasons (it requires division by zero which is not defined).

Numerical Examples of Attenuation-Corrected Correlation and Reliability

Two hypothetical numerical examples illustrate how attenuation correction (10) affects reliability when applied in different estimators: Case 1 represents tests with extreme difficulty level causing radical attenuation in reliability; this case is comparable with the real-life case by Metsämuuronen and Ukkola (2019) discussed above. Case 2 represents tests with incremental difficulty levels in items comparable with the traditions in the achievement testing. The examples are created to highlight the differences between the traditional estimator and the ACERs. Third example comes by a larger simulation of 1440 real-life datasets.

The numerical tables and in-depth discussion in seen in Supplement Appendix 2. Here, only the outlines of the results are discussed. Obviously, in-depth studies are needed to confirm the behavior of the used estimators in real-world datasets as well as in the controlled situations.

Case 1: A Test of Extreme Difficulty Level

Assume a hypothetical dataset of five items with 0–2 points (Tables A1a–A1c in Supplement Appendix 2). This could be a screening test of understanding instructions where all native speakers are expected to get full marks while second language speakers or those with some learning difficulty may make some mistakes in the test items.

Four points highlighted from Case 1 are relevant also later in Case 2. First, using equation (1), the traditional coefficient alpha underestimates the reliability in an obvious manner: $\widehat{\alpha} = 0.352$. The low value is caused by reduction in item variances leading to MEC in observed ρ_{iX}; even at the highest, given the dataset, ρ_{gX} can reach values $\rho_{iX\_Max} = 0.616$–0.894. Equation (11) gives an attenuation-corrected estimate $\widehat{\alpha}_{RAC} = 0.774$. Although the correction of attenuation or deflation in reliability based on the alpha formula is remarkable (0.42 units of reliability), it seems conservative in comparison with the more advanced ACERs: equation (12) gives an estimate $\widehat{\rho}_{TH\_RACX} = 0.834$ and equation (13) $\widehat{\rho}_{\omega\_RACX} = 0.873$; all are notably higher than the traditional alpha and theta ($\widehat{\rho}_{TH\theta} = 0.631$). In comparison with the different ACERs, the original alpha seems deflated by 0.48–0.52 units of reliability. Notably, estimates by omega and maximal reliability cannot be calculated because the correlation matrix is not positively definite.

Second, because some of the MCERs based on changing the entire coefficient have concrete interpretations, their estimates may be valuable benchmarks to the deflation in the traditional alpha. The estimator based on R_{PC} (equation (5)) gives the estimate $\widehat{\alpha}_{RPC} = 0.863$, the estimator based on G (equation (6)) $\widehat{\alpha}_G = 0.902$, and the estimator based on D (equation (8)) gives the value $\widehat{\alpha}_D = 0.886$. The estimates by G and D strictly indicates the proportion of logically ordered test takers in the item after they are ordered by the score; this proportion can be calculated by $p = 0.5 \times G + 0.5$ and $p = 0.5 \times D(g|X) + 0.5$ derived from Metsämuuronen (2021b). For example, by using D, in item g_1 this proportion is $p = 0.5 \times 0.842 + 0.5 = 0.921$, that is, 92.1% of all observations in item g_1 are in a logical order after ordered by the score. In all items together, the average proportion is 92.9%. Hence, the set of items can discriminate very effectively those who got full marks from other test takers. Thus, it seems that attenuation-corrected values reflect more accurately the MEC-free reliability than the original estimate by alpha.

Third, that the magnitude of the estimates by α_{RAC} is lower than the one by α_{RPC} is not a general characteristic. In Case 2, it is to be seen that $\widehat{\alpha}_{RAC} > \widehat{\alpha}_{RPC}$. That the estimate based on G is higher than the one by R_{PC} is not a general characteristic either; it is also a coincidence in the dataset. In real-life settings with two or three categories in the item, G and R_{PC} seem to give estimates that are quite close to each other (see simulations in Metsämuuronen, 2021a, 2022b). However, that the magnitude of the estimates by D are lower than those by G is expected because, with the same pairs of variables, D tends to give more conservative estimates of association than G (see Metsämuuronen, 2021b). Also, that the magnitude of the estimates is higher when using G_2 and D_2 in comparison with G and D is expected because G_2 and D_2 are developed to correct the obvious underestimation obtained by G and D when the number of categories exceeded three (D) or four (G). Using equation (7) based on G_2, we get an estimate $\widehat{\alpha}_{G_2} = 0.910$ and by equation (9) based on D_2, we get $\widehat{\alpha}_{D_2} = 0.894$.

Fourth, to outline, because ρ_{gX} is severely attenuated with items of extreme difficulty level, the estimate of reliability by coefficient alpha of a test with extreme difficulty level is severely attenuated; the traditional alpha may underestimate reliability around 0.42–0.52 units of reliability in comparison with attenuation-corrected coefficients. The simple correction for attenuation (equation (10)) in each item and the related ACERs have a remarkable improvement over the traditional estimator. By comparing these estimates with the other type of MEC-corrected estimates of reliability based on coefficient alpha, we note that the estimate by α_{RAC} seems conservative when the test has an extreme difficulty level. This characteristic is not a general one as is to be seen in Case 2. A simulation regarding the limits and characteristics of ACERs are discussed with Case 3.

Case 2: Incrementally Structured Dataset

Assume a hypothetical dataset as in Tables A2a–A2c (Supplement Appendix 2) of five items with 0 –2 points with incremental difficulty level of items. This could be a short-ish sub-test of "Algebra" as a part of a longer achievement test of mathematics.

Basically, the main result is the same as in Case 1: reliability estimated by the traditional coefficient alpha (($\widehat{\alpha} = 0.411$)), theta ($\widehat{\rho}_{TH} = 0.531$) , and omega $\widehat{\rho}_{\omega} = 0.563$ are deflated because the test comprises both very easy and difficult items; even at the highest in the given dataset, PMC in the extreme items could not exceed values $\rho_{iX\_Max} = 0.451 - 0.482$. The estimates of the ACERs by equations (10)–(12) are $\widehat{\alpha}_{RAC} = 0.790$ $\widehat{\rho}_{TH\_RAC} = 0.838$, $\widehat{\rho}_{\omega\_RAC} = 0.881$, respectively. The estimate by the ACER based on the alpha formula comes quite close to the ones by other types of MEC-corrected estimates by equations (5)–(7) ($\widehat{\alpha}_{RPC} = 0.787$, $\widehat{\alpha}_{G} = 0.785$, and $\widehat{\alpha}_{G_2} = 0.806$, respectively). In the case, the differences between α_{RAC} and other deflation-corrected estimates based on alpha are subtle but the difference between these and the traditional α is notable; the traditional alpha seems to underestimate reliability around 0.38–0.47 units of reliability, traditional theta seems to underestimate reliability around 0.25 units, and traditional omega around 0.24 units of correlation.

Case 3. Larger Simulation of the Behavior of R_{AC} and Attenuation-Corrected Estimator of Reliability

A real-life dataset of 4022 nationally representative test-takers of a mathematics test with 30 binary items (FINEEC, 2018) is used as the "population" in simulation of the behavior of R_{AC} and ACERs in the real-life testing settings. The characteristic of the dataset is discussed in Supplement Appendix 2. The dataset of individual items including several indicators of item–score association is available at http://dx.doi.org/10.13140/RG.2.2.17594.72641. The dataset of reliabilities is available at http://dx.doi.org/10.13140/RG.2.2.27971.94241. The main results of the simulation are collected in what follows.

First note to make is that, with very small sample size ($n = 25$), both Rit and R_{AC} seem to tend to underestimate the population correlation in an obvious manner although R_{AC} less than Rit (Figure A1 in Supplement Appendix 2). Second, except the smallest sample size in the simulation, the R_{AC} in the samples tends to be overestimate the R_{AC} in population mildly with small sample sizes and when the scale in the item is wide. This is understood by the fact that, with small sample sizes, the probability to obtain near-deterministic patterns leading to high magnitudes of R_{AC} is higher than in the (larger) population. With items with a narrow scale ($df(g) < 4$) and with sample sizes around $n = 100$ or higher, the possible overestimation is nominal (see Table A3 in Supplement Appendix 2).

Third, in the simulation dataset, the average estimates of reliability by DCERs are 0.04–0.07 units of reliability higher than those by the traditional estimators using their traditional linking factor and score variable (Table A4 in Supplement Appendix 2). This follows strictly from the fact that magnitude of the estimates by R_{AC} and R_{PC} tend to be higher than of those by *Rit*. That the difference in the magnitude of the estimates in the simulation by the traditional estimators and DCERs is not as dramatic as in Cases 1 and 2 is caused by the fact that the tests in simulation do not allow to prepare tests of extreme difficulty. In the simulation, obtaining tests with extreme difficulty level would have required very short tests using only items with extreme difficulty levels.

Fourth, the estimates using R_{AC} and R_{PC} tend to give estimates with largely the same magnitude (see Figure 1; Figure A2 in Supplement Appendix 2) and systematically higher than those by the traditional estimators. This seems to refer to the phenomenon that both DCERs refer to the *same* latent reliability which is underestimated around 5–8 % by the traditional estimators regardless of the difficulty level of the test items. It seems that, with extreme datasets, the magnitude of the estimates by R_{AC} are mildly higher than those by R_{PC}. This difference is nominal though. Simulation of more extreme datasets would shed more light in this matter.

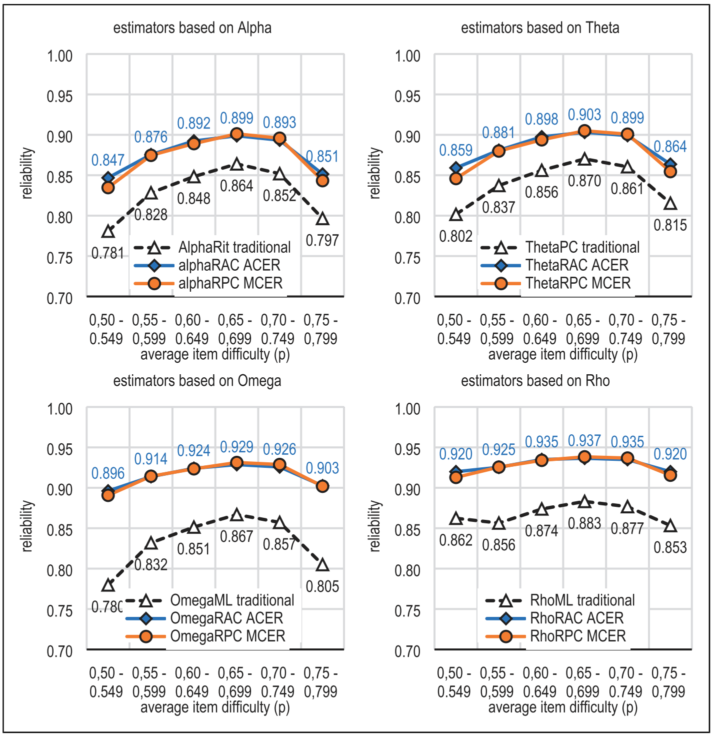

Figure 1. Average estimates of reliability by selected DCERs by the test difficulty.

Fifth, when it comes to the difference between the sample and population estimates, except the smallest sample size, the ACERs give stable estimates; the differences between the sample and population estimates are subtle. As an example, when using alpha as the base and R_{AC} as the linking factor, the deviance between the sample and population estimates is, on average, 0.001–0.003 units of reliability depending on sample size (Table A5 in Supplement Appendix 2). If the sample size is n = 200 (or higher), difference is less than 0.001 units of reliability.

Sixth, because of mild overestimation of population R_{AC}, the ACERs seem to give mild overestimation of the population reliability, specifically, with small sample sizes, polytomous items, and if using rho as the base (Figure A3 in Supplement Appendix 2). The last is expected because rho alone is known to give overestimations with finite sample sizes (Aquirre-Urreta et al., 2019; see also Supplement Figure 3); R_{AC} and rho combined seems to lead to greater overestimation than the other combinations. From this viewpoint, the MCERs based on alpha, theta or omega and using R_{PC} as the linking factor tend to lead to more conservative estimates. However, from the factual estimate viewpoint, R_{AC} and R_{PC} seem to lead largely to estimates of same magnitude (see Figure 1 above).

Seventh, regardless of the width of the scale in the items, the estimates by ACERs seem to bring us nearer the population value than the other estimators except if rho would be selected as the base (Figure A4 in Supplement Appendix 2). When the scale is very wide (more than seven categories), ACERs tend to overestimate reliability mildly, but the values are still nearer the population value in comparison with traditional estimators and MCERs in comparison. Using rho as the base for ACER is not recommendable because of tendency to produce overestimates with finite samples.

Conclusion of Numerical Examples

As a conclusion from Cases 1, 2, and 3, it is known that, first, the proposed R_{AC} and ACERs using R_{AC} as the linking factor may be advantageous in reflecting the true reliability, specifically, when the test includes item with extreme difficulty levels. These types of tests are common in educational settings where the tests are often constructed so that both very easy and very demanding items are included in a test. In these cases, the traditional item–score correlation may be radically attenuated and deflated while R_{AC} gives a plausible alternative to quantify the true association between the item and score variable. Second, a larger simulation based on a real-world dataset suggests that R_{AC} gives estimates that are nearer the population value than PMC with small sample sizes although it seems to overestimate mildly the population R_{AC} when the number of categories in the item exceeds 5. Simulations in this regard would be beneficial. All in all, the larger simulation (Case 3) did not include very extreme datasets. Simulation with datasets of extreme difficulty levels would be beneficial.

Notably, the advance of ACERs in estimating the standard error of the measurement ($S.E.m$) may be notable in the datasets where the item difficulties are extreme leading to an ultimately non-normal score (see Case 1 above). Supplement Appendix 3 shows an example of a comparison of estimates of $S.E.m$ by using traditional estimators and ACERs. It is seen that, because of technical reasons, a magnitude of a rough general estimate of the measurement error may decrease by 36% or more if we use a deflation-corrected alpha instead of the traditional alpha. It seems that, in comparison with other types of DCERs, ACERs combined with R_{AC} give conservative estimate in the case of tests with extreme item difficulty (see Metsämuuronen, 2022a). In any case, selecting wisely estimators of reliability that produce estimates being nearer the true reliability value may give us a notable advance in assessing the accuracy of the test scores.

All in all, ACERs discussed in this article are part of a larger family of deflation-corrected estimators of reliability. Comparing different weight factors used in these estimators as well as comparing ACERs with other DCERs (see Metsämuuronen, 2022a), would be beneficial in finding the best combinations of the base and the weight factor. For instance, some estimators may be more usable or recommendable with binary items and some with polytomous items. Systematic comparisons of different estimators and weight factors in different conditions such as varying the test length, test difficulty, sample sizes, item types, and distributions of the latent variables would be beneficial.

Conclusions and Limitations

Main Results in a Nutshell

The intention in the article was to illustrate the phenomenon of attenuation and deflation in the estimates of reliability and to offer some practical solutions for the problem. The root reason for the attenuation in reliability is the item–score correlation embedded in most of the traditional and widely used estimators of reliability. Because the estimates of correlation by PMC between items and the test score are always attenuated and deflated, this causes attenuation in the estimates of reliability when using the traditional estimators such as coefficient alpha, theta, omega, and maximal reliability. Examples show that, in the extreme cases, the estimates by alpha may be radically deflated to the extent of 0.40–0.50 units of reliability and, in real life-settings, even more than 0.60 units of reliability (see Gadermann et al., 2012; Metsämuuronen & Ukkola, 2019). As a specific solution for the attenuation in the reliability, a new kind of attenuation correction is proposed to replace PMC in the formulas: the attenuation-corrected PMC (R_{AC}) as the proportion of observed correlation of the maximal possible correlation with the given variables. Although the numerical examples in the article were given in the context of measurement modeling, R_{AC} is not restricted to settings related to items and score variable.

Simulations suggest that R_{AC} could be a useful coefficient to describe the association between two variables with scales with a notable difference in width: *RAC* strictly refers to the proportion we obtain of the maximal possible correlation with the given dataset. However, if the sample size is small ($n < 200$), the proportion may be mildly smaller in the population. This is caused by the fact that the probability to obtain deterministic or near-deterministic patterns in a small sample is much higher than in a large population. Deterministic and near-deterministic patterns lead to values $R_{AC}\approx 1$ which are rarely obtained with large sample sizes and wider populations.

The characteristics of the R_{AC} were not studied in the article in-depth; some limits are discussed here. R_{AC} reaches the value 1 when the maximum possible value of PMC is achieved, that is, when the item and the score are in the same order. Value 0 is obtained when the observed correlation is 0. R_{AC} can also reach negative values; because the maximum possible value is always positive the value of R_{AC} is negative when the observed PMC is negative. Hence, R_{AC} reaches the limits of correlation $(-1 \leq R_{AC} \leq +1)$ If, in the further simulations, R_{AC} is found to be an asset in evaluating attenuation in correlations, it may be worth considering reporting routinely as a related statistic to the observed correlation. Specifically, in case the scales differ from each other in an obvious manner as is usual in the measurement modeling settings between item and the score variable, reporting either R_{AC} or the maximal possible correlation given the dataset may help assess the magnitude of possible attenuation or deflation in the observed estimates. Algorithms for estimating the highest possible correlation given the item-and score variance are easy to develop. Maybe, R_{AC} could be considered also when choosing the best correction formula for the r^2 effect sizes (see, e.g., Skidmore & Thompson, 2011; Vacha-Haase & Thompson, 2004).

After the $R_{AC}s$ are calculated, when applied to different base-forms of reliability, attenuation-corrected alpha ($\rho_{\alpha\_RAC\theta}$ or α_{RAC}), theta ($\rho_{TH\_RAC\theta}$), omega ($\rho_{\omega\_RAC\theta}$), and maximal reliability ($\rho_{MAX\_RAC\theta}$) are easy to calculate. These ACERs may remarkably reduce the attenuation in the estimate or reliability—in the numerical examples, deflation in the estimates by coefficient alpha was found to be as high as 0.38–0.52 units of reliability depending on the data structure and the estimator used. In the real-life settings the deflation may be less remarkable; in a larger simulation the deflation was around 0.04–0.07 units of reliability. It can be predicted that when the test includes items with extreme difficulty levels (easy or difficult), ACERs would remarkably correct the attenuation in the estimates of reliability.

Of the initiated ACERs, attenuation-corrected alpha seems to give conservative estimates of the reliability in the case of tests with extreme difficulty level. However, with the tests of incrementally structured difficulty level, the magnitude seems to be at the same level of magnitude as with MEC-corrected estimates of reliability based on R_{PC}, G and G_2. Because ρ_{gX} never reaches the limits of correlation when the scales of the variables are not identical, it is obvious that, in all practical conditions faced in measurement modeling settings, $\alpha_{RAC} > \alpha$, $\rho_{TH\_RAC} > \rho_{TH}$, $\rho_{\omega\_RAC} > \rho_{\omega}$, and $\rho_{MAX\_RAC} > \rho_{MAX}$. Simulations would be beneficial in exploring the behavior of both R_{AC} and different ACERs in different controlled situations. Specifically, simulations with short tests, test with extreme difficulty levels, and studies related to different types of score variables would be beneficial. Obviously, comparisons of different types of DCERs in different conditions would be beneficial. The latter incudes also comparison of Zumbo et al.'s (2007) ordinal alpha and ordinal theta with the DCERs discussed in this article.

Known Limitations of the Study

The study did not discuss varying interpretations and limitations of different coefficients. However, it is known that the estimators based on G and D have concrete interpretations in reflecting the proportion of logically ordered test takers in the dataset (see Metsämuuronen, 2021b). It is also known that estimators based on R_{PC} do not refer to the observed score but something unreachable and theoretical (see the discussion in Chalmers, 2017). However, as an indicator of *theoretical* maximum correlation, R_{PC} could be used as a benchmark to R_{AC}. From this viewpoint, the ACERs (11), (12), (13), and (14) seem to lead us to more practical interpretations of the observed score than those using R_{PC}. Using the latter estimators outside of their original context of principal component- and factor analysis may be debatable though; here, it was assumed that the formulas of theta, omega, and rho *could* be used also as stand-alone estimators without their original contexts.

The study did not tackle the question of possible overestimation of reliability if using attenuation- and MEC-corrected estimators of reliability. However, as a benchmark, if we think that R_{PC} do not *over*estimate the true correlation, it may be relevant to conclude that a MEC-corrected estimator based on R_{PC} such as equation (5) would not overestimate reliability. A relevant question is, what would be the mechanism for overestimation in attenuation-corrected estimator? From this viewpoint, we recall the results by Aquirre-Urreta et al. (2019) that maximal reliability may overestimate the true reliability with finite samples familiar in real-world testing settings. Hence, DCERs based on rho, in general, may tend to overestimate the population reliability with small and smallish sample sizes. Theoretical and empirical studies in the area would be beneficial.

All in all, this article intended to promote discussion of attenuation in reliability and to offer possible practical solutions in the spirit of Schmidt and Hunter (1999) who suggested incorporating the knowledge from attenuation studies to the estimation of measurement error. The closer we can reach the deflation-free estimates of reliability the more accurately we can evaluate the overall quality of the measurement, describe the error in the test scores, correct the estimates in regression and path modeling as well as correct the attenuation in the validity studies and meta-analysis. Hopefully, the attenuation correction in correlation and ACERs proposed in this article are found useful in this endeavor.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jari Metsämuuronen https://orcid.org/0000-0001-6027-0799

Supplemental Material

Supplemental material for this article is available online.

References

Aitken, A. C. (1934). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society*, *4*(2), 106–110. https://doi.org/10.1017/S0013091500008063

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

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

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

Chan, D. (2008). So why ask me? Are self-report data really that bad? In C. E. Lance, & R. J. Vanderberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 309–326). Routledge. https://doi.org/10.4324/9780203867266

Cheng, Y., Yuan, K.-H., & Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educational and Psychological Measurement*, *72*(1), 52–67. https://doi.org/10.1177/0013164411407315

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. https://doi.org/10.20997/sr.19.2.4

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

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).

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*, *17*(3), 1–13. https://doi.org/10.7275/n560-j767

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

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

Greene, V. L., & Carmines, E. G. (1980). Assessing the reliability of linear composites. *Sociological Methodology*, *11*, 160. https://doi.org/10.2307/270862

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

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

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

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

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

Kim, J.-O., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it*. Series: Quantitative applications in the social sciences, no. 13. Sage Publication, Inc.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/001316447003000105

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

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh Section A: Mathematic*, *61*(1), 28–30. https://doi.org/10.1017/S0080454100006385

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

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *1*(1), 98–107. https://doi.org/10.1037/1082-989X.1.1.98

Lord, F. M. (1958). Some relations between Guttman's principal component scale analysis and other psychometric theory. *Psychometrika*, *23*(4), 291–296. https://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.

Martin, W. S. (1973). The Effects of Scaling on the Correlation Coefficient: A Test of Validity. *Journal of Marketing Research*, *10*(3), 316–318. http://dx.doi.org/10.2307/3149702

Martin, W. S. (1978). Effects of Scaling on the Correlation Coefficient: Additional Considerations. *Journal of Marketing Research*, *15*(2), 304–308. https://doi.org/10.1177/002224377801500219

Martinson, E. O., & Hamdan, M. A. (1972). Maximum likelihood and some other asymptotical efficient estimators of correlation in two-way contingency tables. *Journal of Statistical Computation and Simulation*, *1*(1), 45–54. https://doi.org/10.1080/00949657208810003

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

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

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

Metsämuuronen, J. (2017). *Essentials of research methods in human sciences*. SAGE Publications.

Metsämuuronen, J. (2020b). Dimension-corrected Somers' *D* for the item analysis settings. *International Journal of Educational Methodology*, *6*(2), 297–317. https://doi.org/10.12973/ijem.6.2.297

Metsämuuronen, J. (2020a). Somers' *D* as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. *International Journal of Educational Measurement*, *6*(1), 207–221. https://doi.org/10.12973/ijem.6.1.207

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), 283–307. https://doi.org/10.1007/s41237-021-00138-8

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

Metsämuuronen, J. (2022b). The 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., & Ukkola, A. (2019). *Alkumittauksen menetelmällisiä ratkaisuja (Methodological solutions of zero level assessment)*. Publications/Julkaisut 18:2019. Finnish Education Evaluation Centre. https://karvi.fi/app/uploads/2019/08/KARVI_1819.pdf

Moses, T. (2017). A review of developments and applications in item analysis. In R. Bennett, & M. von Davier (Eds.), *Advancing human assessment. The methodological, psychological and policy contributions of ETS* (pp. 19–46). Educational Testing Service. Springer Open. https://doi.org/10.1007/978-3-319-58689-2_2

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurement. *Psychometrika*, *32*(1), 1–13. https://doi.org/10.1007/BF02289400

Olson, U. (1980). Measuring Correlation in Ordered Two-Way Contingency Tables. *Journal of Marketing Research*, *17*(3), 391–394. https://doi.org/10.1177/002224378001700315

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. (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. (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

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modeling approach. *The British Journal of Mathematical and Statistical Psychology*, *57*(Pt 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. https://doi.org/10.1177/0013164417725127

Revelle, W., & Condon, D. M. (2018). *Reliability from α to ω: A tutorial*. https://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 Educational 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. https://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. (1999). Theory testing and measurement error. *Intelligence*, *27*(3), 183–198. https://doi.org/10.1016/S0160-2896(99)00024-0

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

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, *61*(4), 827–868. https://doi.org/10.1111/j.1744-6570.2008.00132.x

Skidmore, S. T., & Thompson, B. (2011). Choosing the best correction formula for the Pearson r2 effect size. *The Journal of Experimental Education*, *79*(3), 257–278. https://doi.org/10.1080/00220973.2010.484437

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

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

Thorndike, R. L. (1949). *Personnel selection*. Wiley.

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

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*(4), 473–481. https://doi.org/10.1037/0022-0167.51.4.473

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika*, *42*(4), 579–591. https://doi.org/10.1007/BF02295980

Yang, H. (2010). Factor loadings. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 480–483). SAGE Publications. https://doi.org/10.4135/9781412961288.n309

Zaionts, C. (2022). Real statics using excel. Polychoric correlation using solver. http://www.real-statistics.com/correlation/polychoric-correlation/polychoric-correlation-using-solver/(Accessed 02 03 2022).

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. https://doi.org/10.22237/jmasm/1177992180