Practical Assessment, Research, and Evaluation

Volume 27 Article 10

2022

How to obtain the most error-free estimate of reliability? Eight Sources of Deflation in the Estimates of Reliability to Avoid

Jari Metsämuuronen
Finnish National Education Evaluation Centre (FINEEC)

Recommended Citation

Metsämuuronen, Jari (2022) "How to obtain the most error-free estimate of reliability? Eight Sources of Deflation in the Estimates of Reliability to Avoid," *Practical Assessment, Research, and Evaluation*: Vol. 27, Article 10.
DOI: https://doi.org/10.7275/7nkb-j673
Available at: https://scholarworks.umass.edu/pare/vol27/iss1/10

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 10, June 2022 ISSN 1531-7714

How to obtain the most error-free estimate of reliability? Eight Sources of Deflation in the Estimates of Reliability to Avoid

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

The reliability of a test score is usually underestimated and the deflation may be profound, 0.40 - 0.60 units of reliability or 46 - 71%. Eight root sources of the deflation are discussed and quantified by a simulation with 1,440 real-world datasets: (1) errors in the measurement modelling, (2) inefficiency in the estimator of reliability within the selected measurement model, (3) inefficiency in forming of the score variable (X) as the manifestation of the latent trait θ, (4) non-optimal characteristics of the items (g_i) in relation to the estimator, and (5) inefficient weight factor, that is, coefficient correlation (w_i) that links θ with the observed values of the test item (x_i), (6) a small sample size, (7) extreme test difficulty, and (8) a narrow scale in the score. If willing to maximize the probability that the estimate of reliability would be as close as possible the true, population value, these sources should be avoided, or their effect should be corrected by using deflation-corrected estimators of reliability.

Keywords: Reliability; Attenuation in reliability; Attenuation in correlation; Item–total correlation; Coefficient alpha; Coefficient theta, Coefficient omega; Maximal reliability

1. Introduction

Traditionally, the concept of reliability is used to quantify the amount of random measurement error that exists in a score variable generated by a compilation of multiple test items. However, in the large-scale testing settings such as in the national level assessments as well as in international inquiries 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 (SE) in different parts of the ability scale (see, e.g., Foy & LaRoche, 2019). Nevertheless, even if the average random error is less accurate than the one obtained by more complex strategies, it may still serve as a rough indicator of SE of the score.

Traditionally, an estimate of reliability (*REL*) serves the researcher in several ways. First, it is used in quantifying the amount of average random error in a score variable, that is, the standard error of measurement (*S.E.m*):

$$S.E.m. = \sigma_E = \sigma_X \sqrt{1 - REL}$$

derived strictly from the basic definition of reliability

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

(Gulliksen, 1950), where σ_X^2, σ_T^2, and σ_E^2 refer to the variances of the observed score variable (X) and the unobserved true score (T) and error (E) related to the classic relation of $X = T + E$. Second, an estimate of reliability of the total score serves many ways in further use of the score: in assessing the (overall) quality of the

measurement (e.g., Gulliksen, 1950; Metsämuuronen, 2017, 2022b), 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 purposes, we want to obtain as accurate an estimate of reliability as possible.

Guttman (1945), in his seminal article, showed that the observed estimates of reliability tend to be underestimates of the true, population reliability. This fundamental finding strictly concerns such classical estimators of reliability as Brown–Spearman prediction formula for parallel partitions of a test (ρ_{BS}; Brown, 1910; Spearman, 1910), Flanagan–Rulon prediction formula for non-parallel partition of a test (ρ_{FR}; Rulon, 1939), and the coefficient called the greatest lower bound reliability (ρ_{GLB}; Jackson & Agunwamba, 1977; Woodhouse & Jackson, 1977; Ten Berge & Zegers, 1978 onwards) because these all are special cases of Guttman's (1945) coefficient λ_4, which was shown to underestimate reliability. Revelle and Zinbarg (2009) note that ρ_{GLB} does not necessarily lead to the lowest *bound* but just a lower *value* in comparison with coefficient omega total (ρ_{ω}; Heise and Bohrnstedt, 1970; McDonald, 1970 onwards) which is known to give lower estimates than the coefficient rho or maximal reliability (ρ_{MAX}; see, e.g., Li, 1997; Raykov, 1997b; 2004; see also Cheng, Yuan, & Liu, 2012). Hence, it seems that also coefficient omega underestimates reliability. However, recall the result of Aquirre-Urreta, Rönkkö, and McIntosh (2019) that ρ_{MAX} may give *over*estimates with finite sample sizes. Hence, assessing the amount of underestimation is not always easy.

Another generally known outcome of Guttman's article is that the most widely used estimator of reliability, the coefficient alpha (ρ_{α}; Kuder & Richardson, 1937; Jackson & Ferguson, 1941), known also as Cronbach's alpha (Cronbach, 1951), underestimates reliability because it equals Guttman's coefficient λ_3 also shown to underestimate reliability. Guttman himself condensed his results as follows: "Reliability has often been underestimated by the conventional formula […]. Many tests are more reliable than they have been considered to be" (Guttman, 1945, p. 260). It is generally accepted that if the measurement errors of individual items correlate as is the case when, for instance, using the same stimulus for several items, alpha may overestimate reliability (see e.g., Trizano-Hermosilla & Alvarado, 2016). However, the inflation may be nominal in the whole test when comparing it with radical *deflation* in alpha (up to 0.60–0.70 units of reliability, see Gadermann, Guhn, & Zumbo, 2012; Metsämuuronen, 2022a, 2022b) caused by technical or mechanical underestimation of correlation embedded in the estimators of reliability.

Based on literature, the root causes for the underestimation of reliability can be divided into five categories: (1) errors in the measurement modelling, (2) inefficiency in the estimator of reliability within the selected measurement model, (3) inefficiency in forming of the score variable (X) as the manifestation of the latent trait θ, (4) non-optimal characteristics of the items (g_i) in relation to the estimator, and (5) inefficient weight factor, that is, coefficient correlation (w_i) that links θ and the observed values of g (x_i). These are discussed in this article, first, by referring to relevant literature and, second, using an empirical dataset of 1,440 tests.

When it comes to underestimation of reliability, two terms are in use: 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 of the estimation (see the discussion of the terms in, e.g., Chan, 2008; Gadermann, Guhn, & Zumbo, 2012; Metsämuuronen, 2022a, 2022b; Revelle & Condon, 2018). Deflation is closer the focus in this article, and it is connected to another concept called here "mechanical error in estimates of correlation" (MEC; see, e.g., Metsämuuronen, 2021a, 2022a), that is, a characteristic of estimators of correlation to underestimate the true correlation because of technical or mechanical reasons. The practicalities related to deflation and MEC are discussed later in Sections 2.5 and 2.6, and practical implications related to deflation will be discussed in the Discussion (Sections 6.2 and 6.3).

2. Root causes for the deflation in reliability

Deflation in the estimates of reliability may be radical. With certain types of datasets, typically with very easy, very demanding, and tests with incremental

difficulty levels in items common in educational assessment, the estimates by ϱ_{α} and ϱ_{MAX} are found to have been deflated notably: ϱ_{α} up to 0.70 units of reliability and ϱ_{MAX} over 0.40 units or 46%–71% (see examples in, for instance, Gadermann et al., 2012; Metsämuuronen, 2022b, 2022c; Metsämuuronen & Ukkola, 2019; Zumbo, Gadermann, & Zeisser, 2007). Most probably the same phenomenon concerns also estimates by ϱ_{TH} and ϱ_{ω}. Reasons behind deflation of this size can be found in five directions. These are discussed in what follows.

2.1 Reasons of the underestimation embedded in the measurement model

Assume a general, simplified, one-latent variable measurement model combining the observed values of an item $g_i(x_i)$, a latent variable (θ), and a weight factor w_i that links θ with x_i:

$$x_i = w_i\theta + e_i, \quad (1)$$

generalized from the traditional model related to the practicalities of factor analysis (for instance, McDonalds, 1999; Cheng et al., 2012). This model generalizes to the score variable as a compilation of items ($X = \sum_{i=1}^{k} x_i$) as

$$X = T + E$$
$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} w_i\theta + \sum_{i=1}^{k} e_i \quad (2)$$

(see, Metsämuuronen, 2022a, 2022b, 2022c). In the general model, the theoretical, unobservable θ may be manifested as a varying type of relevantly formed compilation of items including a raw score (θ_{RAW}), principal component score variable (θ_{PC}), factor score variable (θ_{FA}), theta score formed by the item response theory (IRT) or Rasch modelling (θ_{IRT}), or a nonlinear compilation of various kinds (θ_{NonL}). The general weight factor w_i may be either a coefficient of correlation or the factor- or principal component loadings (λ_i) ranging $-1 \le w_i \le +1$. Different options of coefficient of correlation are compared by Metsämuuronen (2022a)—these are discussed later in Section 2.6.

The general model includes the root causes for the deflation in the estimates of reliability: on the top of the model itself and the estimators of reliability, the term $w_i\theta$ in Eqs. (1) and (2) refers to a fact that the estimates of reliability vary depending on the manifestation of the latent variable θ, characteristics of the item i, and the weight factor w. Then, if we use improper measurement model not fitting the dataset, inefficient estimator of reliability, inefficient score variable, inefficient scales in item, and inefficient estimator of correlation as the weighting factor, the estimate of reliability may be far below the true value. These five elements are discussed in what follows. The effect of the model itself in the underestimation is discussed in Section 2.2 and of the estimators of reliability in Section 2.3. Effect of the score variable is discussed in Section 2.4. The effect of item characteristics such as item difficulty and the number of categories is discussed in Section 2.5 and the effect of the weight factor in Section 2.6.

2.2 Errors in the measurement modelling causing deflation in reliability

Traditionally, measurement models related to latent variable are divided into three: models where the test partitions (including sub-tests and single items in a compilation of a test) are either parallel, tau-equivalent, or congeneric (e.g., Lord, Novick, & Birnbaum, 1968). In the strictest and oldest models based on parallelism we assume that statistical characteristics in the partitions g and h are parallel, that is, the true values of the same test-taker are identical in each partition ($T_g = T_h$), leading to the realization that correlations between the partitions, if being more than two, are identical, which leads us to assume unidimensionality in the phenomenon when single items are taken as partitions. This also leads to assume that the correlations between the partitions and the score variable are identical ($w_i = w_j = w$) as well as are the measurement errors ($e_i = e_j = e$). Also, the classical test theory assumes that measurement errors are uncorrelated, that is, the test items should be independent from each other.

Some generally known estimators of reliability based on this model are ϱ_{BS} and Kuder and Richardson (1937) formula 21 (ϱ_{KR21}) discussed above. Because ϱ_{BS} is a special case of Guttman's (1945) λ_4, and coefficient λ_4 was shown to underestimate reliability "*no matter how the test is split*" (Guttman, 1945, p. 260, emphasis original), the estimates by ϱ_{BS} are always underestimates of the population reliability.

The measurement model based on (essential) tau-equivalency loses the strict assumptions of parallelism

to some extent. This makes sense because the assumptions in the models based on parallelism are rather restricting and difficult to meet in real-life testing settings. In tau-equivalent models we assume that the true values of the same test-taker are (essentially) identical in the partitions ($T_g = T_h$) but the partitions need not be parallel in the strict sense although the sub-tests or scales in the items should be equally long; notably, if using unidentical scales in items leads us to a violation of tau-equivalency. These assumptions lead us to assume that the weights are equal ($w_i = w_j = w$), indicating unidimensionality in the phenomenon, but the measurement errors need not be equal ($e_i \neq e_j$) although they should not correlate with each other.

Some known estimators within the tau-equivalency are ϱ_{FR} for non-parallel partitions with equal lengths, Gutman's λ_3 and λ_4, Kuder and Richardson (1937) formula 20 (ϱ_{KR20}) for non-parallel binary items, and coefficient alpha (ϱ_α) for polytomous items with identical scales. Because λ_3 and λ_4 were shown to underestimate the population reliability (Guttman, 1945) and the other estimators are special cases of those, all these estimators tend to underestimate population reliability. Traditionally, the attenuation in ϱ_α has been connected to such errors related to the measurement modelling as violations in tau-equivalency, unidimensionality, and uncorrelated errors (see the discussion around alpha in, for instance, Davenport et al., 2015; 2016; Green & Yang, 2009, 2015; McNeish, 2017; Novick & Lewis, 1967; Raykov & Marcoulides, 2017; Trizano-Hermosilla & Alvarado, 2016). The approximations of the underestimation in estimates related to this kind of modelling error have varied from nominal (Raykov, 1997a) up to 11% (Green & Yang, 2009).

The least restricted family of measurement models is based on congeneric partitions. In these models, the true values of the same test-taker need not be identical in the partitions ($T_g \neq T_h$), leading to lose the assumption of equally long sub-tests or partitions or of the same scale in items. Also, weights need not be equal ($w_i \neq w_j$), allowing multidimensionality in the phenomenon, the measurement errors need not be equal ($e_i \neq e_j$), and they need not be independent from each other; the last can be modelled during the estimation of reliability.

Estimators of reliability based on the congeneric partitions are many. For two sub-scores—as alternatives for ϱ_{BS} and ϱ_{FR}—we have coefficients by Horst (ϱ_H; Horst, 1951), Angoff and Feldt (ϱ_{AF}; Angoff, 1953; Feldt, 1975), and Raju (ϱ_β; Raju, 1977). If the partitions are equally long, the magnitude of these estimates gets the relation $\rho_{FR} = \rho_\beta \leq \rho_{SB} = \rho_H \leq \rho_{AF}$ (Warrens, 2016), that is, the Angoff–Feldt coefficient would give the highest estimate, and if the variances of the partitions are equal, $\rho_{FR} = \rho_{SB} = \rho_{AF} \leq \rho_H = \rho_\beta$ (Warrens, 2016), that is, Horst's coefficient and Raju's β give the highest estimate. Consequently, from the underestimation viewpoint, the other estimates underestimate the reliability more than these if the conditions relevant for ρ_H and ρ_β or ρ_{AF} are met.

For the case that we are interested in using items with different scales in the estimation and willing to use the raw score of the items, the congeneric alternative for coefficient alpha would be Gilmer–Feldt coefficient (ϱ_{GF}; Gilmer & Feldt, 1983) also known as Feldt–Raju coefficient (e.g., Feldt & Brennan, 1989) or as Feldt–Gilmer coefficient (e.g., Kim & Feldt, 2010). This estimator tends to give higher values than alpha.

For the case we want to work with weighted scales within the framework of factor analysis, three estimators are in a more common use: ϱ_ω known also as McDonald's omega total, ϱ_{MAX} known also as composite reliability or Raykov's rho (Raykov, 1997b) or Hancock's H (Hancock & Mueller, 2001), and coefficient theta (ϱ_{TH}; chronologically, Lord, 1958; Kaiser & Caffrey, 1965; Armor, 1973) based on principal component analysis, known also as Armor's theta. It is known that ϱ_{TH} maximizes ϱ_α (Greene & Carmines, 1980), Bentler's alphamax or alpha-O maximizes ϱ_{TH} (Bentler, 1968; Greene & Carmines, 1980), and the estimates by ϱ_{MAX} are higher than those by ϱ_ω (see, e.g., Cheng et al. 2012). Then, of these four estimators, ϱ_{MAX} is known to give the highest estimates, and ϱ_{TH}, and ϱ_ω give estimates with higher magnitude than ϱ_α if the loadings are not equal. Hence, if ϱ_{MAX} is taken as a benchmark, both ϱ_α, ϱ_{TH}, and ϱ_ω seems to underestimate reliability assuming that the conditions optimal for ϱ_{MAX} such as large sample size are met. Empirical section studies, among others, what the

effect of sample size and characteristics of the test is in deflation.

2.3 The estimator of reliability causing deflation in reliability

Above, it was noted that even if being consistent within a measurement model, we have several estimators that produce slightly different estimates of the same latent reliability of which some are more deflated than the others. Differences between the estimators are obvious when we compare the estimators based on different partitions of the test: selecting the partitions with lower correlation we get estimate with a lower magnitude of reliability than if we select partitions with a higher correlation; this is the whole idea of ϱ_{GLB} (Guttman, 1945). From this viewpoint, ϱ_{BS} is based on *strictly parallel* partitions, ϱ_{FR} is based on *non-parallel* partitions with equal length, ϱ_{α} is based in partitions with *average correlation* between the partitions, and ϱ_{GLB} is based on selected the partition with the *highest correlation* between the partitions (see Revelle & Condon, 2018). Such estimators as Revelle's β (Revelle, 1979; see also Zinbarg, Revelle, Yove, & Li, 2005) and McDonald's hierarchical omega (McDonald, 1999) are based on selecting the partition with the *lowest correlation* and, hence, these could be called the estimators of the lowest lower bound (ϱ_{LLB}) of reliability. Knowing that both ϱ_{BS}, ϱ_{FR}, ϱ_{α}, and ϱ_{GLB} underestimate reliability, estimators in the family of ϱ_{LLB} give *obvious* underestimations of the population reliability.

Usually, comparing such widely used estimators as ϱ_{α}, ϱ_{TH}, ϱ_{ω}, and ϱ_{MAX} (see later Eqs. 3–6) does not make sense because, except ϱ_{ω} and ϱ_{MAX}, the manifestation of θ differs estimator-wise and, hence, the differences may be caused by this (see Section 2.4) rather than the estimator itself. However, of ϱ_{ω} and ϱ_{MAX}, as using the same score and the same maximum likelihood (ML) estimation (Jöreskog, 1967 onwards), we know that the estimates by ϱ_{MAX} are higher than those by ϱ_{ω}; hence, the formula of ϱ_{ω} seems less effective than the formula of ϱ_{MAX} if the practical requirements for ML-estimation such as large sample size are fulfilled.

Comparing ϱ_{α} and ϱ_{MAX} or ϱ_{α} and ϱ_{ω} as estimators is less clear because of the different manifestation of the latent variable and the weighting factor. This area is largely unstudied. Some light is shed on this in the empirical section. Another related aspect to the discussion is that ϱ_{MAX} is known to give *over*estimates with finite or small sample sizes (Aquirre-Urreta, Rönkkö, & McIntosh, 2019) as discussed above. With very small sample sizes, the risk for deterministic or near-deterministic patterns in the dataset increases. With these patterns of item discrimination, neither of ϱ_{ω} and ϱ_{MAX} can be used because no factor solution is given. In the empirical section, this phenomenon is studied in real-life settings with finite sample sizes.

2.4 Inefficiency in forming of the score variable causing deflation in reliability

The effect of the estimator itself is sometimes difficult to separate from the effect of the score variable; after all, we tend to use different estimators with different types of scores. It is known that ϱ_{TH} maximizes ϱ_{α}. This may be at least partly caused by the fact that ϱ_{TH} is based on more efficient (weighted) compilation of the items than ϱ_{α}. It is generally known that the raw score used in alpha formula as the manifestation of the latent variable (θ_X) is not as efficient in discriminating the test-takers as the optimal linear compilation or weighted compilation of the items would be (see, e.g., Li, 1997). Early contributions of seeking the "optimal linear compilation" of the items can be traced to Lord (1958), Stouffer (1950), Guttman (1941), and Thompson (1940). Later, the expressions were unified for maximal reliability by Li (1997).

Weighting the items have led in three main approaches of the manifestation of the latent variable θ: principal component scores (θ_{PC}), factor scores (θ_{FA}), and theta scores by Rasch- and item response theory (IRT) models (θ_{IRT}); the last is, factually, a special case of factor score variable though. Of the many estimation methods related to factor analysis, the one based on ML estimation (MLE) is known to produce the maximal estimates for factor loadings. This leads to maximal estimate of reliability, and, then, using *other* estimation methods would lead us, consequently, to underestimate reliability. MLE embeds two specialties related to estimation of reliability. First, estimates by MLE cannot be calculated for only two variables, that is, if we have genuinely two items in the test (see discussion in Bridgeman, 2016) or we interpret that split-halves are two "items" with a wide scale, MLE cannot be used to produce the factor loading for these "items". Some other methods such as principal axis factoring (PAF), however, could be used. Second, estimates by MLE are not necessarily stable with small

sample sizes because of possible deterministic or near-deterministic conditions in any of the test items; the first leads to no factor solution and the latter to (artificially) high estimates. The empirical section provides further information regarding this too. The pure effect of the manifestation of the latent variable is not necessarily easy to assess unambiguously because the estimators themselves including the weight factor w_i differ from each other. Some light is shed on this in the empirical section.

2.5 Characteristics of the items causing deflation in reliability

Single items are the basis of the test score. Traditionally, the items are divided into objective ones such as multiple choice- or short answer type of questions and subjective ones such as productive items in mathematics or essay type questions in subjects related to humanities and natural sciences which require subjective evaluation to form the score (e.g., Mehrens & Lehmann, 1991; see also Bridgeman, 2016; Metsämuuronen, 2017). The scales of the test items are usually non-continuous and ordinal. In achievement testing, we tend to use binary (0 = incorrect; 1 = correct) or slightly graded scales, and, in attitude scales, we often use such ordinal scale as 4 to 5-point Likert scales. Even the advanced routines of measurement modelling including IRT modelling are based on these conventions. The estimators of reliability are not restricted to these forms of scales, but the routines related to item writing, item scoring, and item analysis often are. Two characteristics of the item are raised here as noteworthy when it comes to deflation in the estimates of reliability: item difficulty and the scale of the item.

The item difficulty is one of the clearest sources of mechanical error in the estimates of correlation causing deflation in the estimates by product-moment correlation coefficient (PMC; Pearson 1986; see, e.g., Metsämuuronen, 2021a, 2022a) embedded in the most widely used estimators of reliability including ϱ_α, ϱ_{TH}, ϱ_ω, and ϱ_{MAX} in the form of item–score correlation (*Rit*) or principal component- or factor loading (λ_i) (see the formulae and literature in Section 2.6). When the item difficulty is extreme—either extremely easy or extremely difficult—the loss of information by PMC approximates 100%: the more extreme is the difficulty level the lower is the maximal possible value achieved by PMC irrespective of the fact that the latent correlation would be perfectly $\varrho = 1$. Hence, *Rit* and λ_i are always underestimates of the true association between an item and the latent variable. This is the technical reason why the estimates of reliability in the empirical datasets may have been radically deflated (up to over 70%; see Gadermann et al., 2012; Metsämuuronen, 2021a, 2022b, 2022c; Metsämuuronen & Ukkola, 2019) (see Section 2.6).

The scale of item is strictly related to the deflation in reliability: the less categories in an item, the more deflation in the estimates of item–total correlation which is inherited to the estimates of reliability (see Metsämuuronen, 2021a, 2022a). The pure effect of item scale in the empirical datasets seems largely unknown. The empirical section sheds light on this too.

2.6 Inefficiency in coefficients of correlation causing deflation in reliability

Above, four sources of underestimation in reliability are seen to cause deflation in the estimate of reliability to a certain extent although their effect, in many practical testing settings, may be small. Far more grave deflation in reliability have been obtained in empirical studies related to the selection of the weighting factor w_i already discussed above with item difficulty. As noted above, with certain types of datasets, typically with very easy, very demanding, and tests with incremental difficulty levels in items common in educational assessment, the estimates by ϱ_α and ϱ_{MAX} are found to have been deflated notably: ϱ_α up to 0.70 units of reliability and ϱ_{MAX} over 0.40 units or 46%–71% as discussed above. Deflation of this size is no more of a matter of just modelling error; it is remarkable and worth studying. The empirical section discusses this issue.

Gadermann and colleagues (2012), Zumbo and colleagues (2007), and Metsämuuronen (2016, 2020a, 2021a, 2021b; 2022a) argue that the reason for the radical deflation in the estimates of reliability has to do with PMC. PMC is embedded in the most widely used formulae including ϱ_α, ϱ_{TH}, ϱ_ω, and ϱ_{MAX} in the form of item–score correlation (*Rit*) or principal component- or factor loading (λ_i). In the formula of alpha, PMC is seen strictly as $Rit = \rho_{iX}$:

$$\rho_{\alpha} = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right) = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times \rho_{iX}\right)^2}\right) \quad (3)$$

(Lord et al., 1968) because the estimated population variance (σ_X^2) can be expressed by item variances (σ_i^2) and item–score correlation (ρ_{iX}), that is,

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

where k is the number of items in the compilation. Also, we remember that the principal component- and factor loadings are, essentially, PMCs between an item and a score variable (e.g., Yang, 2010). Then, in coefficient theta, PMC is seen as the principal component loading (λ_i):

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

In coefficient omega, PMC is seen as the factor loading (λ_i):

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

as well as in rho:

$$\rho_{MAX} = \frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(\lambda_i^2/\left(1-\lambda_i^2\right)\right)}}. \quad (6)$$

The reason for the deflation in the estimates of reliability is that the estimates by PMC are known to be deflated due several sources of MEC when the scales of two variables differ from each other. This is always is the case with a score variable (X) and test item (g_i) (see algebraic reasons, for instance, in Metsämuuronen, 2016, 2017; and simulations, for instance, in Martin, 1973; 1978; Olsson, 1980; Metsämuuronen, 2021a, 2022a). PMC is, specifically, affected by the item difficulty. When item variance σ_i^2 approximates 0, that is, when the proportion of 1s (p) or 0s ($1–p$) in the binary case approximates 0, PMC approximates 0 irrespective of the latent, true correlation. These kinds of sources of MEC are many including item difficulty, number of categories in the scale, and number of tied cases in the dataset. Metsämuuronen (2021a, 2022a), for instance, noted seven such sources of which all cause negative bias in the estimates by PMC.

According to simulations (Metsämuuronen, 2021a, 2022a), some good options of the weight factor w_i are polychoric correlation (R_{PC}; Pearson, 1900, 1913), bi- and polyreg coefficient (R_{REG}; see Livinstone & Dorans, 2004; Moses, 2017), Goodman–Kruskal gamma (G; Goodman & Kruskal, 1954), dimension-corrected G (G_2; Metsämuuronen, 2021a), and attenuation-corrected PMC and eta (R_{AC}, E_{AC}; Metsämuuronen, 2022d). These estimators are, practically speaking, free of MEC when it comes to reflect the true, perfect correlation. Quite good options although not as good as those above would be Somers delta (D; Somers, 1962; see Metsämuuronen, 2020a) and dimension-corrected D (D_2; Metsämuuronen, 2020b; corrected in 2021a); these are affected by the number of tied cases. Consistently with the idea of the general measurement model, the weight factor may vary item-wise even within a test; some estimators may be more efficient with binary items (e.g., G and D) although some others may be more efficient with items with both binary and polytomous scales (e.g., R_{PC}, G_2, and D_2).

By replacing *Rit* and λ_i in Eqs. (3) to (6) with a totally different coefficient being less affected by MEC leads us to estimators of reliability called "MEC-corrected estimators of reliability" (MCER, Metsämuuronen, 2022b) or, if attenuation-corrected estimators are used, to "attenuation-corrected estimators of reliability" (ACER; Metsämuuronen, 2022c). In what follows, these both are called by a common name "deflation-corrected estimators of reliability" (DCER; Metsämuuronen, 2022a, 2022b; 2022c). Notably, Zumbo's and colleagues' (2007)

ordinal alpha and ordinal theta also belong to the extended family of DCERs; instead of simply changing the factor loading itself, the estimation starts by replacing the matrix of inter-item PMCs by a matrix of R_{PC}s.

As suggested by Metsämuuronen (2021a, 2021b, 2022a, 2022b, 2022c), the general (theoretical) bases for DCERs could be based on coefficient alpha (Eq. 3):

$$\rho_{\alpha\_wi\theta} = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times w_{i\theta}\right)^2}\right), \quad (7)$$

coefficient theta (Eq. 4):

$$\rho_{TH\_wi\theta} = \frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}w_{i\theta}^2}\right), \quad (8)$$

coefficient omega total (Eq. 5):

$$\rho_{\omega\_wi\theta} = \frac{\left(\sum_{i=1}^{k}w_{i\theta}\right)^2}{\left(\sum_{i=1}^{k}w_{i\theta}\right)^2 + \sum_{g=1}^{k}\left(1-w_{i\theta}^2\right)}, \quad (9)$$

and coefficient rho (Eq. 6):

$$\rho_{MAX\_wi\theta} = \frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(w_{i\theta}^2/\left(1-w_{i\theta}^2\right)\right)}}. \quad (10)$$

Using theta, omega, and rho outside their original context of principal component- and factor analysis may be debatable; this issue is discussed later. As discussed above, the term $w_{i\theta}$ refers to the fact that the estimate differs depending on the selected coefficient of correlation w, characteristics of an item i, and the manifestation of the latent variable θ. By wisely selecting the weighting factor, it is possible to remarkably reduce the deflation in reliability. This also solves largely the issue related to the item effect; as examples, such estimator as G with binary items, and R_{PC} and G_2 with binary and polytomous items all are, practically speaking, MEC-free in many conditions such as difficulty level, number of categories in the scale, and number of tied cases in the dataset discussed above (see Metsämuuronen, 2021a, 2022a). Of these, R_{PC} may lead to a theoretical reliability because the coefficient itself does not refer to the observed score but to a latent, unobservable score (see the critique in Chalmers, 2017).

To outline the discussion so far, the reasons of deflation in reliability can be traced to, at least, five sources which all may cause simultaneously deflation in the estimates of reliability. However, studies of the phenomenon tend to be fragmentary and some areas may be even unstudied. The empirical section explores the effect of simultaneous sources in real-life settings.

3. Research questions

It is reasonable to think that the five sources of deflation in reliability may all have an effect at the same time in a cumulative manner. There also may be more sources of deflation than what discussed above. Their common effect is largely unknown under different simultaneous conditions. The empirical section studies the behaviour of the four widely used estimators ϱ_{α}, ϱ_{TH}, ϱ_{ω}, and ϱ_{MAX} in different conditions related to real-life testing setting. The specific research questions are: (1) What is the magnitude of the effect of sample size, number of categories in the scales of the score and items, as well as of difficulty level of the test in deflation of reliability and (2) How effectively the estimators reflect the population estimate; to what extent the estimators under- or overestimate the population reliability in real-life testing settings with finite or small sample sizes.

4. Methodology

4.1 Measurement model and estimators used in the empirical section

The general measurement model discussed in Section 2.1 is applied in the empirical section. Mainly the traditional estimators ϱ_α, ϱ_{TH}, ϱ_ω, and ϱ_{MAX} (Eqs. 3–6) with their intended original score variables (θ_X for alpha, θ_{PC} for theta, and θ_{FA} with MLE for omega and rho) and weight factor (*Rit* for alpha and λ_i for theta, omega, and rho) are in focus. If only two items (of wide scales) form the score, PAF is used to estimate the factor loadings. Some benchmarking comparisons are made by using DCERs (Eqs. 7–10) and using R_{PC} and G_2 as the linking factor.

4.2 Datasets and tests used in the study

A real-world representative national-level dataset of 4,022 test-takers of a mathematics test with 30 binary items (FINEEC, 2018) is used as the "population". In the original dataset, $\varrho_\alpha = 0.885$, $\varrho_{TH} = 0.890$, $\varrho_\omega = 0.887$, and $\varrho_{MAX} = 0.895$, item discrimination ranged $0.333 < Rit < 0.627$ with the average $\overline{Rit} = 0.481$, and the difficulty levels of the items ranged $0.24 < p < 0.95$ with the average $\bar{p} = 0.63$

Ten random samples with $n = 25$, 50, 100, and 200 test-takers in each were drawn from the original dataset, imitating different sizes of finite sample sizes typical in real-life testing settings, ranging from a typical classroom testing ($n = 25$) to a test for large student group ($n = 200$). In each of the 10×4 datasets, 36 shorter tests were produced by varying the number of items, difficulty levels of the items, and the length of the scale of the item ($df(g)$ = number of categories in the scale – 1), and in the score ($df(X)$ = number of categories in the scale – 1). The polytomous items were constructed as partitions of the original binary items. As a result, the datasets[1] in simulation consisted of 14,880 partly related test items from 1,440 tests with a varying number of test-takers ($n = 25$, 50, 100, and 200) and items ($k = 2$–30, $\bar{k}$=10.33, std. dev. 8.621), lower bound of reliabilities ($\varrho_\alpha = 0.55$–0.93, $\overline{\rho_\alpha}$=0.850, std. dev. 0.049), the average difficulty levels ($\bar{p}$=0.50-0.76, $\bar{\bar{p}}$=0.66, std. dev. 0.052), and width of the scales in the items ($df(g) = 1$–14, $\overline{df(g)}$=4.57, std. dev. 3.480) and in the score ($df(X) = 10$–27, $\overline{df(X)}$=18.06, std. dev. 3.908).

5. Results

5.1 Effect of the sample size and number of items in the compilation

The estimates by ϱ_α, ϱ_{TH}, ϱ_ω, and ϱ_{MAX} in the 1,440 tests are compared with each other as well as with the known "population". The first note to make is that the datasets with the smallest sample size produce remarkable amount of deterministic or near deterministic patterns (13–16% of the estimates with $n = 25$; Table 1) where the estimates by omega and rho are not defined, or rho produced reliability of $\varrho_{MAX} \approx 1$ even if only one of the factor loadings appeared to be very near the value 1. This phenomenon could be connected with an alternative concept of reliability, "sufficiency of information" by Smith (2005): It seems that the small sample sizes do not give sufficient amount of information for ML-estimates to produce credible estimates of factor loadings for credible estimates of reliability by omega and rho.

Second, of the four estimators in comparison, the estimates tend to be the highest by rho (average 0.875) and the lowest by alpha (average 0.850); this is expected because of their known behaviour. The difference between the estimates gets smaller by the sample size; although the average difference with the smallest sample size ($n = 25$) is 6.4% (= (0.871−0.815)/0.871), it is only 1% with the highest sample size ($n = 200$). The magnitude of the estimates by theta tend to be slightly higher (average 0.858) than those by omega

[1] 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. The dataset of individual items (n = 14,880) including several 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.

Table 1. Basic statistics of the estimators in comparison

Sample Size	Mean				N				Std. Deviation			
N	Alpha	Theta	Omega	Rho	Alpha	Theta	Omega	Rho	Alpha	Theta	Omega	Rho
25	0.815	0.835	0.819	0.871	360	360	314	304	0.073	0.062	0.075	0.059
50	0.859	0.866	0.862	0.881	360	360	360	360	0.037	0.036	0.037	0.035
100	0.863	0.867	0.865	0.877	360	360	360	360	0.025	0.025	0.025	0.025
200	0.864	0.866	0.865	0.872	360	360	360	360	0.022	0.021	0.021	0.021
Total	0.850	0.858	0.854	0.875	1440	1440	1394	1384	0.049	0.042	0.047	0.037

(0.854) although not exceeding those by rho. Third, all estimates give both over- and underestimates in comparison with the population estimate, specifically, with very small sample sizes (Figure 1).

Rho differs from the others in its tendency to overestimate slightly the reliability with all sample sizes (up to 1.1%) although the magnitude with $n = 200$ is very small (0.001 units of reliability or 0.11%) (Figure 2). This result confirms the warning by Aquirre-Urreta and colleagues (2019) that rho tend to give overestimates with finite samples. Unlike the other estimators, rho tends to overestimate reliability irrespective of the length of the test (indicated by the number of items in the compilation, k; see Figure 2). Notably, estimates by rho with $k = 2$ are not overestimated; these are based on factor loadings by PAF instead of ML. Not only rho overestimates reliability with small sample sizes, also its behaviour is radically more unpredictable in comparison of omega (see Figure 3). Of the conservative estimators, estimates by theta tend to be slightly closer to the population reliability than those by alpha and omega.

5.2 Effect of the number of categories in items and scores

In the simulation, the scales of the scores were kept reasonably wide (maximal points 20–30). However, even though the highest score could be 20 or 30, not all values of the potential scale were actualized; with small sample sizes the variety of different values is smaller than with larger sample sizes leading to $df(X) \geq 10$.

All estimators produce remarkable underestimates when the scale of the score is narrow ($df(X) < 15$) although, with test score of wider scale (($df(X) > 15$), all estimators tend to produce estimates close the population value (Figure 4). In alpha and omega, the deflation may be up to 0.11–0.12 units of reliability (or 14–15%) and, with theta, around 0.09 units (or 11%). A possible confounding factor is that the tests with the narrowest scales ($df(X) = 10$–12) were also the most extremes ones compiled of the smallest sample size ($n = 25$) with the most difficult set of items (see Section 5.3). The systematic nature in the phenomenon indicates though that the effect is related primarily to the scale and not the sample size. Hence, in practical settings, we may expect underestimation, specifically, with tests with a narrow scale and when the tests are extreme in their difficulty level. The latter is a known characteristic of tests (see Section 2.6). Studies related to very short test in this regard would be beneficial.

Notably, unlike the scale of the score, the scale of the *item* does not explain the underestimation in real-life datasets. Rho slightly overestimates the reliability except when the scale of items exceeds six. A possible confounding factor at the range of $df(g) > 7$ (not seen in Figure 4) is that the tests with this wide item scale tended to be short in terms of number of items in the compilation ($k = 2$–3). Then, with $k = 2$, PAF-estimate was calculated instead of ML-estimate. Again, of the conservative estimators, estimates by theta underestimate reliability slightly less than those by alpha and omega.

5.3 Effect of the difficulty level of the items

The test items formed eight sets with different difficulty levels. The easiest tests were compiled with the easiest items and their partitions and, in these, the maximum possible score was 24 and 26 points. The most difficult tests were the shortest ones with 20 to 22 points maximum compiled of the most demanding items and their partitions. By random sampling, some tests appeared to be more extreme than the others. However, tests with extreme difficulty levels were not obtained, and difficult tests (average $0.50 < p < 0.55$) and easy tests (average $0.75 < p < 0.80$) are rare in the datasets, 2.6% and 1.1%, respectively and, hence, their low number may not allow generalization.

Figure 1. Under- and overestimation by the estimators

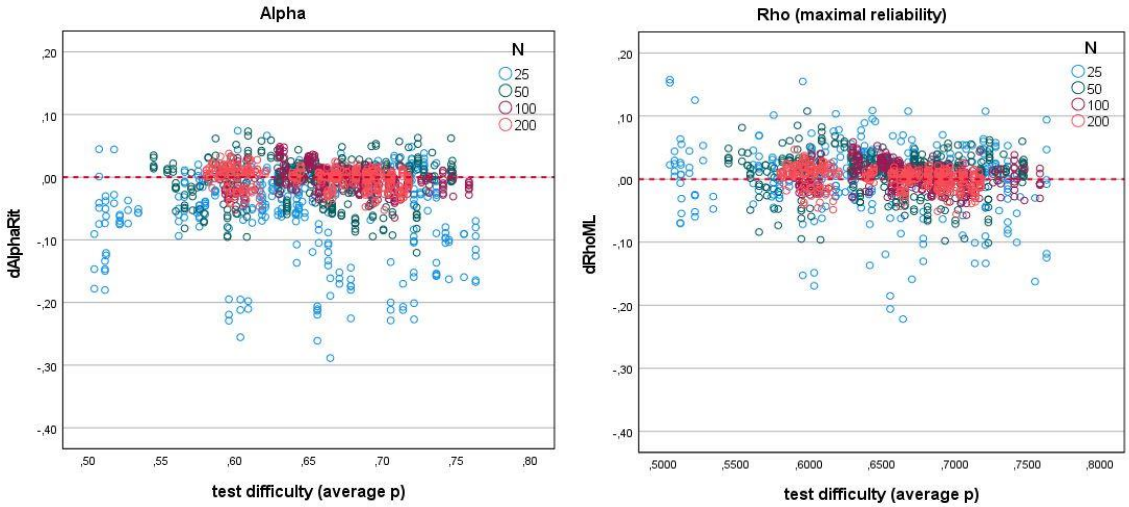


Figure 2. Average under- and overestimation by the estimators by sample size and number of items

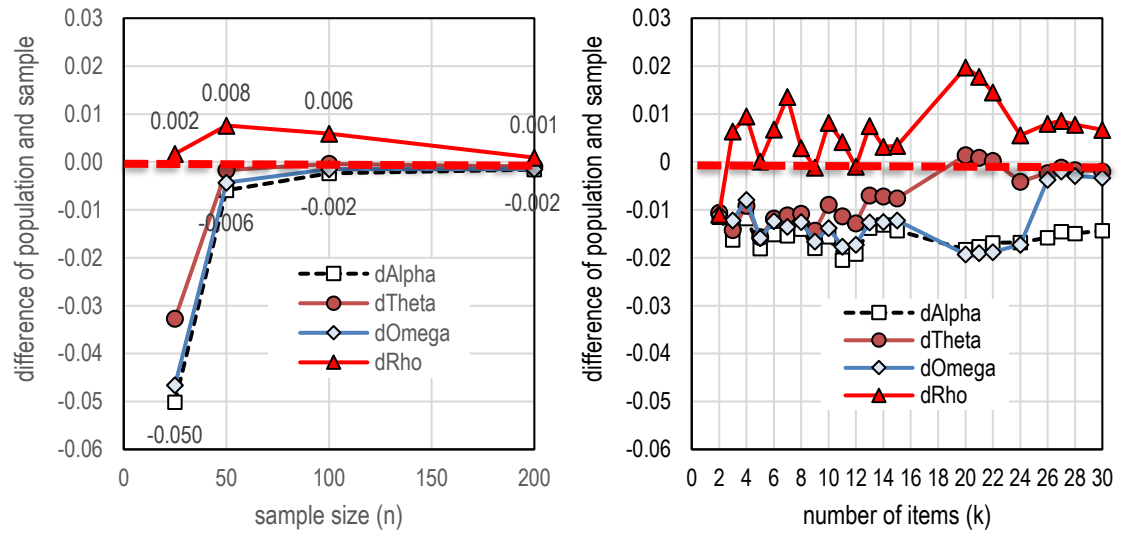


Figure 3. Relation of alpha, omega, and rho

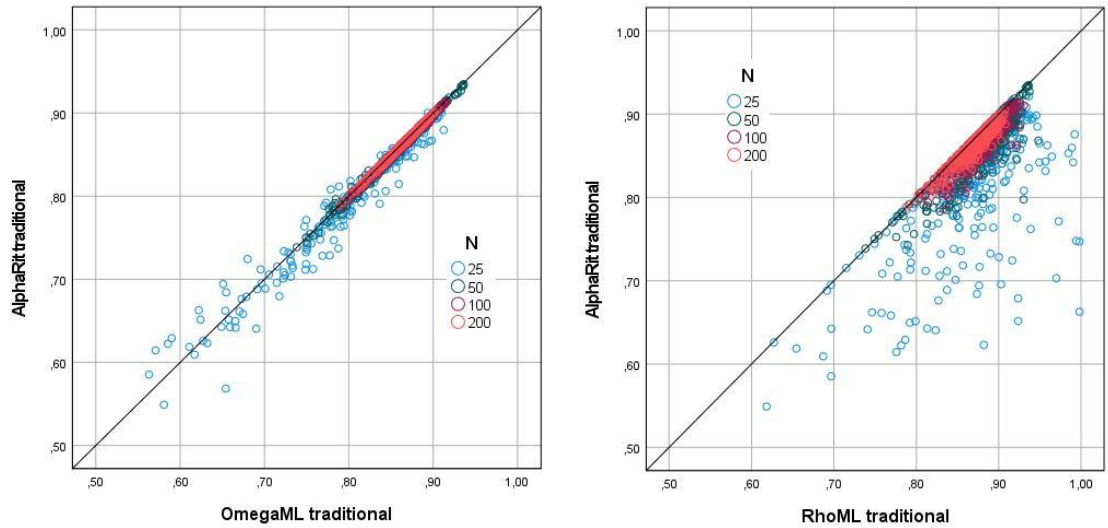
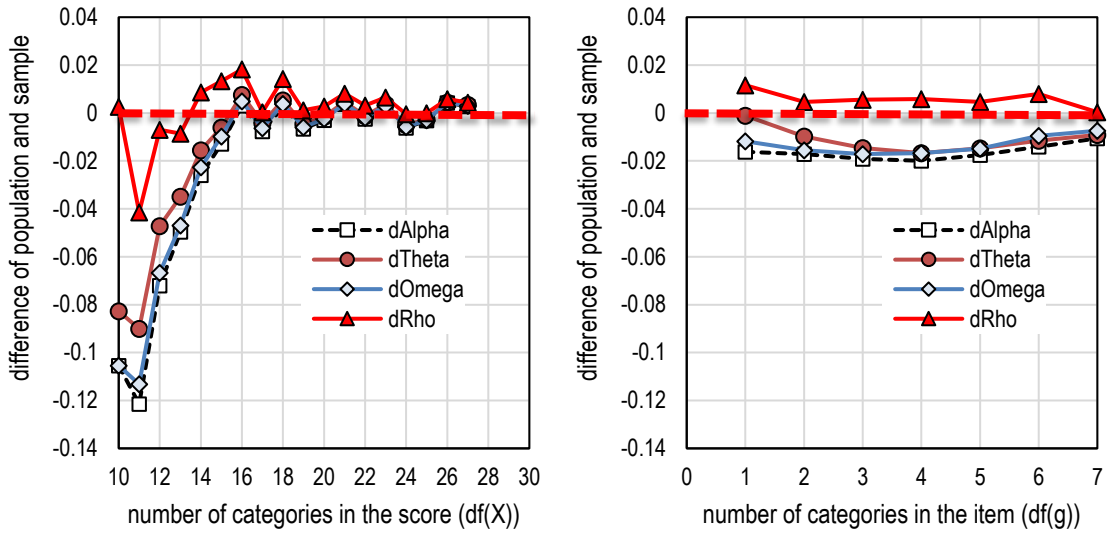


Figure 4. Under- and overestimation by the estimators by df(X) and df(g)



All the estimators in the most extreme datasets tend to give estimates that are far off the population parameter (Figure 5). The estimators by alpha and omega are underestimated by 0.06–0.07 units of reliability (7–8%). Notably, the behaviour of rho differs from the others: it tends to overestimate the reliability of the difficult tests and underestimate that of easy tests. Again, of the conservative estimators, estimates by theta underestimate reliability slightly less than those by alpha and omega. Notably, also, based on empirical findings, we would expect to see far more grave deviation from the population value with the extreme difficulty levels of the test.

Notably, these results have relevance when it comes to the conditional *S.E.m*, that is, random error at different parts of the ability scale. It seems that the traditional estimators or reliability are prone to give notable underestimation in both extremes of ability scale. Deflated reliabilities lead to artificially high standard errors. It seems that DCERs are more stable in this respect (see later Figure 8). This matter is elaborated in section 6.2 with a more extreme dataset.

5.4 Effect of the estimator

Above, it was discussed that assessing the pure effect of the estimator itself is sometimes difficult because both the score variable and the weight factor w_i may vary. Here, this is studied by comparing all four estimators by harmonizing θ and w_i. The raw score is used as the manifestation of the latent variable and R_{PC} as the weight factor.

The estimators based on rho overestimate slightly the population reliability with small sample sizes (Figure 6). Nevertheless, if we compare the estimates by other estimators with the ones by rho, the pure effect of the estimator seems the most notable with tests with a small number of categories in the score (6–8% between ϱ_α and ϱ_{MAX}), difficult or easy items (5%), and the smallest sample sizes (4%). Notably, omega seems to benefit more than theta and alpha in changing the weight factor: the estimates are notably closer the population value if used a deflation-corrected estimator in the estimation than by using factor loading (cl. Sections 6.1–6.3; see also Section 6.5).

5.5 Effect of the selection of the weigh factor

Selection of the weight factor leads us to the extended family of deflation-corrected estimators of reliability. In what follows, two alternative estimators or correlation, R_{PC} and G_2, are used as examples of behavior of DCERs in relation of the traditional estimators. With these DCERs, the raw score was used as the manifestation of θ instead of the traditional factor score. If the factor score variables were used as the manifestation of θ, the outcomes between the estimators may have been slightly different.

Using theta, omega, and rho outside of their traditional context is, undoubtedly, debatable. However, we may think that the estimates by using R_{PC} and G_2 instead of the traditional λ_i are outcomes of renewed procedures on principal component- and factor analysis where the factor loadings are R_{PC} and G_2

Figure 5. Under- and overestimation by the estimators by the test difficulty

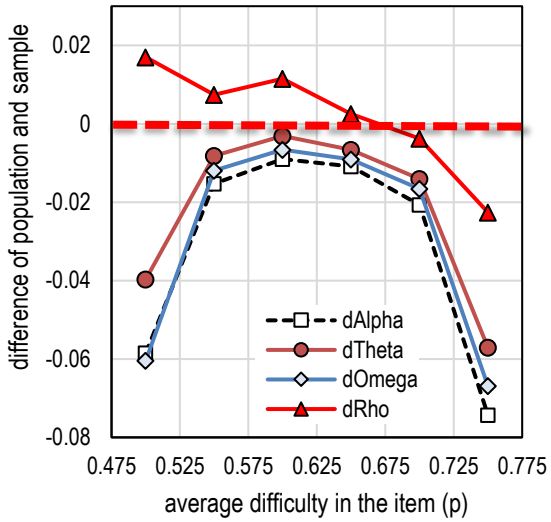
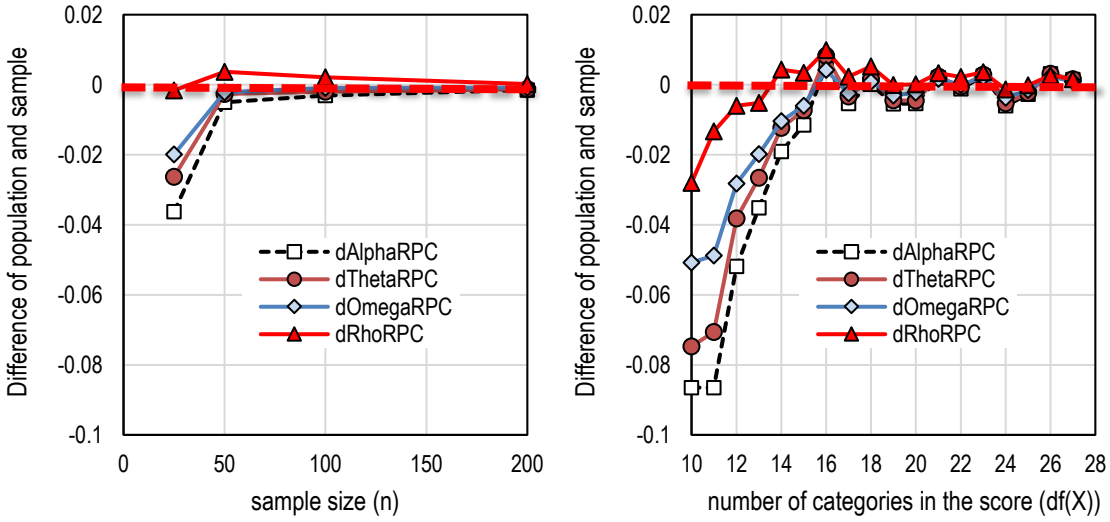
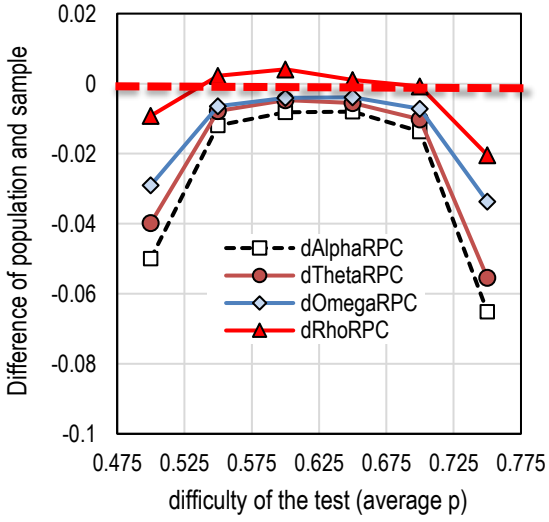


Figure 6. Effect of the estimator after harmonized the score variable and weight factor





instead of PMC (cl. ordinal theta). The rationale in using R_{PC} and G_2 (or some other relevant option) is that they are practically MEC-free unlike *Rit* and λ_i both of which are sensitive to item difficulty. Of the alternatives, R_{PC} leads us to the theoretical reliability because it refers to unobservable score with no strict relevance with the observed score (see Chalmers, 2017) as discussed above. G_2 leads to a more practical interpretation of reliability because the embedded coefficient *G* has a strict interpretation to refer to the proportion of logically ordered test-takers in the whole set of items after they are ordered by the score (see Metsämuuronen, 2021b).

The main advance of DCERs seems to come with the small sample sizes (Figure 7). This is seen the clearest in the estimators based on omega (Figure 8). Although, with the lowest sample size (n = 25), the traditional omega underestimates the population reliability by 0.047 units of reliability or 5.4%, DCERs underestimate half of this, 0.020 or 2.2% (using R_{PC}) and 0.022 units or reliability or 2.4% (using G_2). Specifically, the advantage of DCERs is seen if the scale of the score variable is narrow ($df(X) < 14$) and with very difficult and very easy tests. As an example, when the scale of the score is $df(X) = 11$, the traditional omega underestimates the population reliability by 0.11

Figure 7. Traditional estimators and DCERs by the sample size

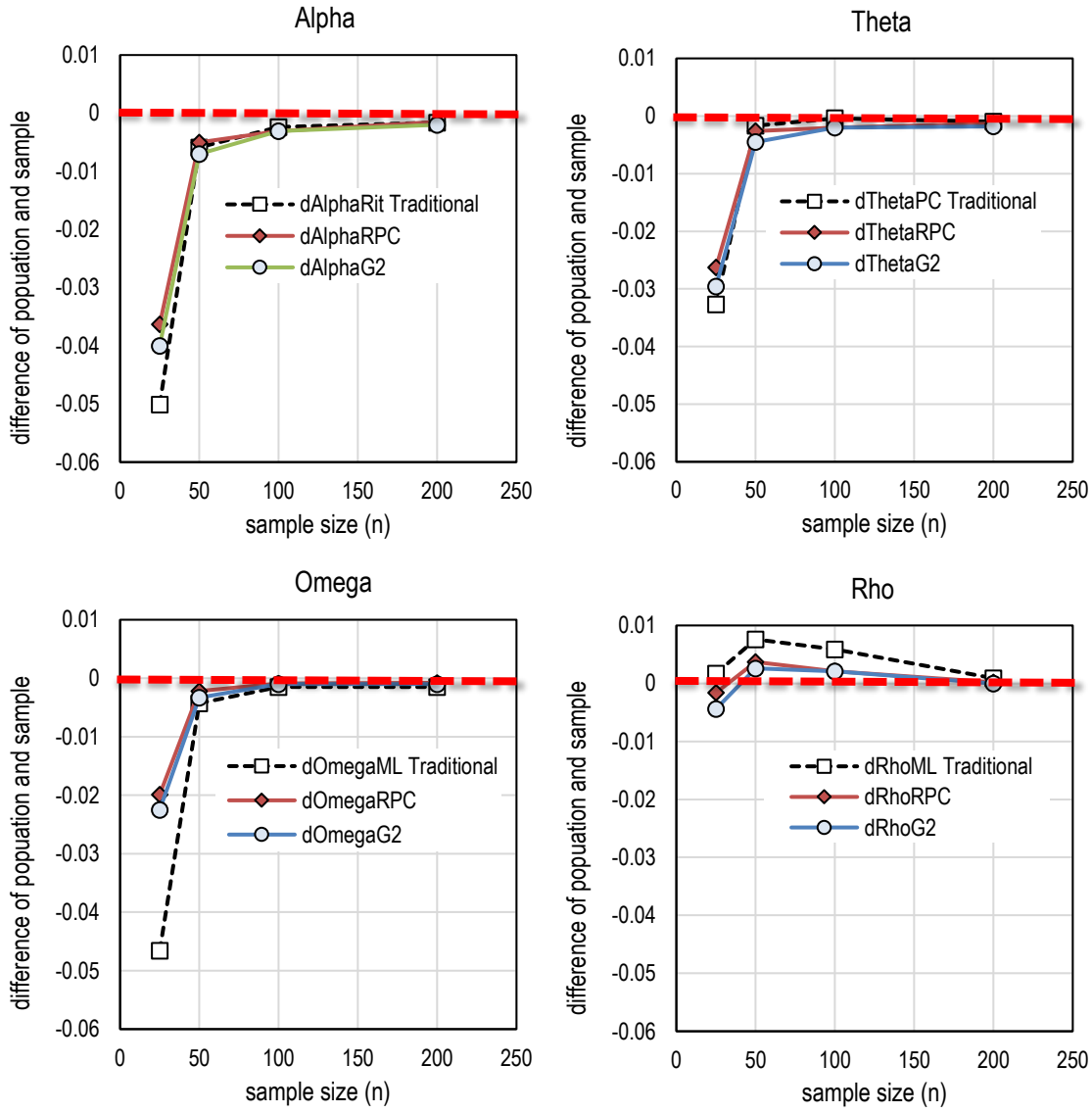


Figure 8. Difference of omega and DCERs by scale length and item difficulty

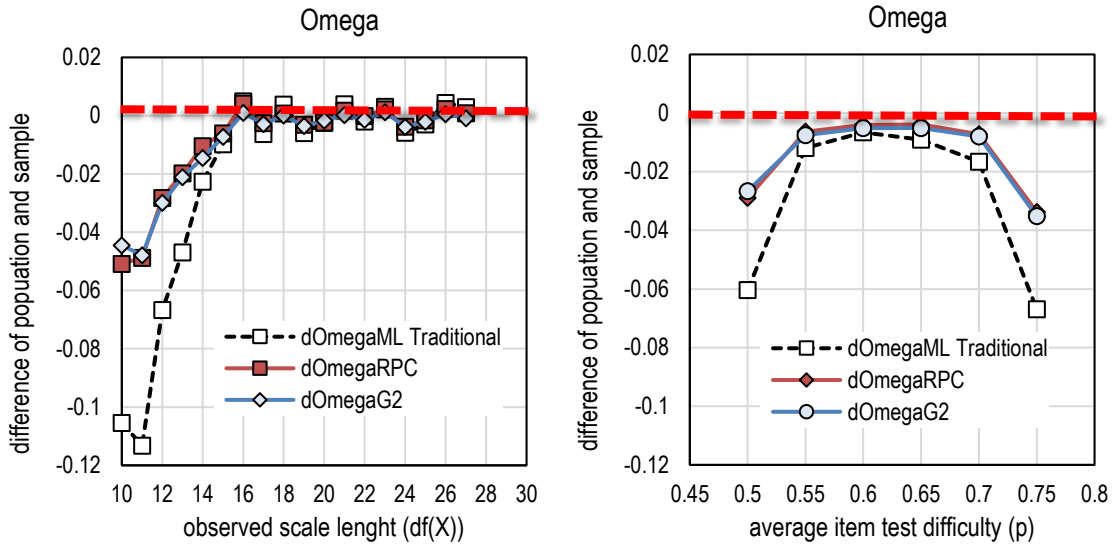
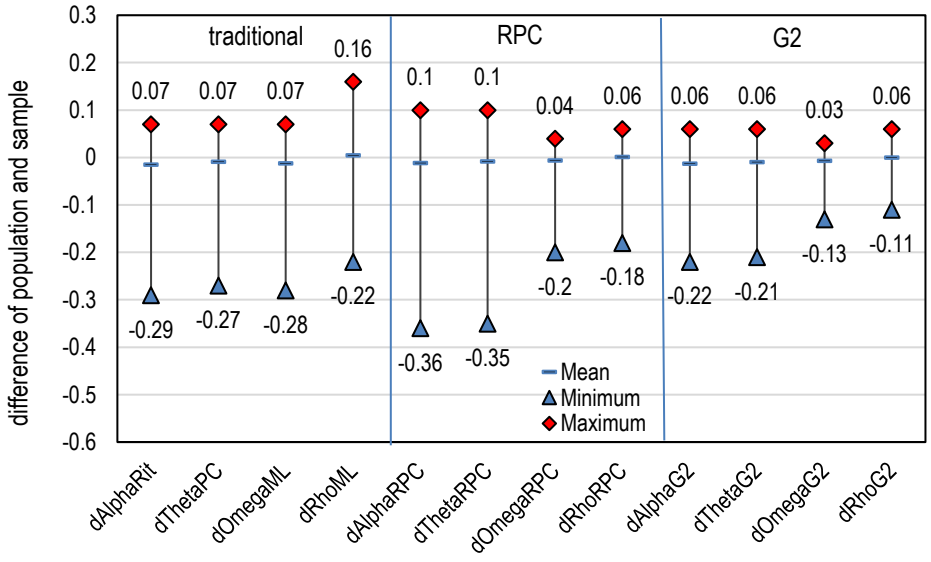


Figure 9. Range of difference between population and sample in different estimators



units of reliability or 13.5% while both DCERs underestimate just 0.05 units or reliability or 5.5%, that is, almost one third less than the traditional omega. With difficult and easy tests, the average advance by using R_{PC} or G_2 is 3% although, by the simulation, we do not know what the difference would be with extremely difficult and easy tests. Estimates based on rho are more stable in comparison with the others (see Figures 1 and 6), but they tend to overestimate slightly the population reliability. From this viewpoint, DCERs bring some advantage: estimates by DCERs tend to be *less* overestimated than those by the traditional rho (see Figure 8). Hence, it seems that the DCERs based on R_{PC} and G_2 give us more stable estimates than the traditional estimates. This is seen also in the narrower range in estimates (Figure 9): As an example, while the range in the difference between the sample and population estimates by the traditional coefficient alpha is 0.36 units of reliability, by using G_2 instead of *Rit* in the formula of alpha, the range in the dataset is narrowed to 0.28 units of reliability (22%).

6. Conclusions and limitations

6.1 Conclusions

The starting point of the article was the generally known characteristic of reliability to be underestimated when using certain estimators; the estimates in the empirical datasets may be, sometimes, radically deflated. Several causes for the deflation were discussed. To summarize the effects of these sources, according to the literature, the violations against the measurement model may have a nominal (1%) to remarkable (11%) effect in deflation. In the simulation, the pure effect of the estimator was found to be up to 8% with tests with a small number of categories in the score and 5% with somewhat difficult and easy tests, and 4% with very small sample sizes. The effect of inefficient scales of items was found to be near zero in the real-life settings although the effect is seen in the theoretical datasets (see Metsämuuronen, 2021a). The effect of the test length or the small number of categories in the score seems notable: up to 15% if $df(X) = 10–12$—shorter tests were not included in the simulation. Small sample size may influence 3–6% to the deflation, and difficulty level of the test 7–8%, except with very extreme difficulty levels of the score with which the deflation may exceed 70% if the weight factor is selected inefficiently (see Section 6.2)—these kinds of tests were not included in the simulation. All these conditions causing deflation may occur at the same time.

As general notes of the four estimators in the study, first, rho tends to give overestimates with small sample sizes up to $n = 200$—higher sample sizes were not used in the simulation. With very small sample sizes ($n < 50$), it is also prone to fail to give the solution because of the random deterministic patterns in the sample. Hence, it is not recommended to use maximal reliability with small sample sizes. From this viewpoint using omega would be a better option as it tends to give more conservative estimates than rho. However, with very small sample sizes, omega used in its original context of factor analysis with ML-estimation also may fail to give a solution.

Second, theta appeared to be surprisingly good option as such in many conditions studied in the simulation. It gives conservatives estimates, that is, it tends to give underestimates, but the deflation in the estimates is smaller than it is in alpha and omega. Even with small sample sizes the estimates by theta are closer to the population value than those by any of the other estimators in comparison. Theta is vulnerable to a narrow scale in the score and the extreme difficulty level, but not that much as are alpha and omega.

Third, both omega and rho could benefit from changing the traditional ML-estimate of factor loading to some other coefficient which would be less affected by MEC than PMC. In practical words, if we use rho or omega as the base and R_{PC} or G_2 (or some other deflation-corrected estimator of correlation) as the weighting factor instead of the traditional factor loading (PMC), both deflation-corrected rho and omega seems to tend to give estimates that are notably closer the true, population value than the estimates by the traditional estimator. This is true, specifically, with omega: notable advance would be gained with tests of extreme difficulty level, with a narrow scale in the score, and with small sample sizes. Unlike omega and rho, theta does not seem to benefit from the replacing the principal component score by R_{PC} and G_2. Another question is whether the estimate by DCER is, factually, an overestimation. Based on the results from the simulation that both alpha, theta and rho tend to underestimate reliability even if Rit and λ_i is changed to a better-behaving coefficient, this is unlikely; what would be the mechanism for the overestimation?

6.2 Practical example of the effect of DCERs in conditional standard errors

To give a practical example of how the deflation in reliability affects the conditional standard errors in the extreme of ability scale, the extremely easy dataset (n = 7,770) by Metsämuuronen (2022b; 2022c; originally in Metsämuuronen and Ukkola, 2019) discussed in Section 2.6 is re-analyzed. Originally, the test was a screening test of proficiency in the language used in the factual test; only the test-takers with second language status (L2) were expected to make mistakes in the test items. For the reanalysis, we may think that the eight items in the test represent a part of a test forming the lower part of the ability scale. Alternatively, these items could be taken as the easiest items in the adaptive testing; how accurately can these items discriminate between the lowest scoring test-takers from the other? The advance of DCERs over the traditional estimators may be notable in these kinds of datasets where the item difficulties are extreme leading to an ultimately non-normal score (see

Metsämuuronen, 2022b). Descriptive statistics of the dataset are collected in Tables 2a and 2b.

Let us assume a setting related to adaptive testing so that the eight items represent a set of items given to a screening test for further sets of items. From the second items onwards, we start to estimate reliabilities and related standard errors of the score. As the estimators of reliability, the traditional theta (Eq. 4) and omega (Eq. 5) with their original weight factors (principal component and factor loadings λ_i by MLE[2]) are compared with DCERs based on theta and omega using *G* as the weight factor (Eqs. 8 and 9). Usually, in complex settings, the standard errors are estimated using complex strategies (see, e.g., Foy & LaRoche, 2019). Here the traditional estimate is calculated ($S.E.m = \sigma_X \sqrt{1-REL}$). Table 3a collects the information related to the score variables. Tables 3b, 3c, and 3d show the estimates of weight variables; 3b shows the principal components loadings for the traditional theta, 3c shows the factor loadings for the traditional omega, and 3d shows the estimates of correlation between items and the score by *G* for the DCERs. In the DCERs, the score variable is the raw score although the result would be identical if the score formed by IRT modelling would have been used (see Metsämuuronen, 2022b). The reason for this is that, when only one test version is in use, the *order* of the test-takers *is identical* irrespective of using the raw score or IRT score. Table 3e collects the estimates of reliability and standard errors in each step adding one item to the test.

Table 2a. Descriptive statistics of the test items from Metsämuuronen & Ukkola, 2019 (N = 7,770)

Item (g)	Range	Mean	p	Std. Deviation	Variance
g1	0−1	0.96	0.96	0.186	0.0348
g2	0−1	0.98	0.98	0.126	0.0160
g3	0−1	0.99	0.99	0.088	0.0078
g4	0−1	0.91	0.91	0.287	0.0824
g5	0−2	1.78	0.89	0.610	0.3715
g6	0−1	0.98	0.98	0.122	0.0150
g7	0−2	1.97	0.985	0.211	0.0446
g8	0−2	1.98	0.99	0.169	0.0285

Table 2b. Descriptive statistics of the score from Metsämuuronen & Ukkola, 2019 (N = 7,770)

Score	freq.	%	IRT Theta	SE(th)	bias
0	0	0	-6.02	1.996	0.533
1	0	0	-4.616	1.066	0.071
2	0	0	-3.793	0.832	0.005
3	4	0.1	-3.241	0.7	-0.012
4	7	0.1	-2.853	0.627	-0.011
5	6	0.1	-2.534	0.595	-0.003
6	20	0.3	-2.232	0.594	0.006
7	42	0.5	-1.906	0.622	0.008
8	146	1.8	-1.499	0.672	-0.007
9	822	10.4	-1.01	0.729	-0.043
10	926	11.6	-0.532	0.844	-0.111
11	5904	75.2	0.093	1.334	-0.350

[2] For only two items (SUM1-2), principal axis factoring (PAF) is used.

Table 3a. Statistics related to the scores with gradually increasing length

	score						
	SUM1-2	SUM1-3	SUM1-4	SUM1-5	SUM1-6	SUM1-7	SUM1-8
number of items (k)	2	3	4	5	6	7	8
Score range	0−2	0−3	0−4	0−6	0−7	2−9	3−11
Mean $\bar{X}$	1.948	2.940	3.849	5.634	6.619	8.592	10.573
Variance σ_X^2	0.053	0.068	0.177	0.580	0.612	0.683	0.765
Std. Dev. σ_X	0.231	0.261	0.421	0.762	0.782	0.826	0.875

Table 3b. λ_i Principal component loadings related to PC scores for Theta

	SUM1-2	SUM1-3	SUM1-4	SUM1-5	SUM1-6	SUM1-7	SUM1-8
g1	0,725	0,567	0,598	0,583	0,524	0,496	0,447
g2	0,725	0,608	0,483	0,478	0,516	0,502	0,430
g3		0,723	0,624	0,616	0,593	0,564	0,605
g4			0,586	0,569	0,506	0,500	0,468
g5				0,266	0,260	0,254	0,204
g6					0,429	0,448	0,375
g7						0,321	0,288
g8							0,633

Table 3c. Factor loadings (MLE) related to factor scores for Omega

	SUM1-2	SUM1-3	SUM1-4	SUM1-5	SUM1-6	SUM1-7	SUM1-8
g1	0.226	0.215	0.357	0.352	0.314	0.305	0.276
g2	0.226	0.241	0.243	0.250	0.302	0.305	0.260
g3		0.576	0.372	0.379	0.395	0.378	0.471
g4			0.340	0.332	0.297	0.304	0.291
g5				0.121	0.129	0.132	0.111
g6					0.231	0.251	0.213
g7						0.165	0.160
g8							0.512

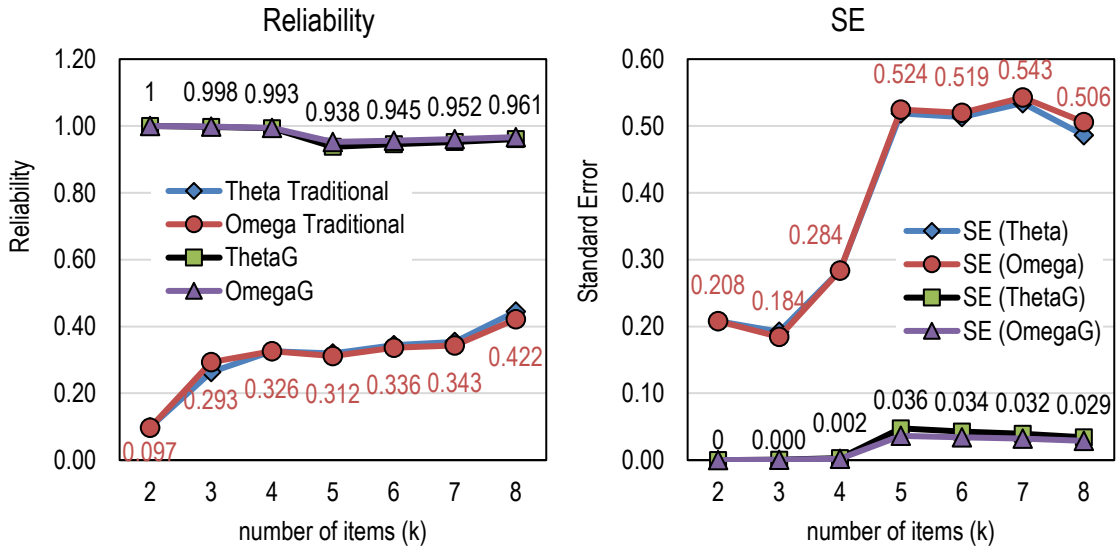
Table 3d. Goodman–Kruskal G_{iX} for DCERs

	SUM1-2	SUM1-3	SUM1-4	SUM1-5	SUM1-6	SUM1-7	SUM1-8
g1	1	0.998	0.993	0.870	0.869	0.858	0.857
g2	1	0.997	0.980	0.851	0.858	0.849	0.846
g3		0.998	0.989	0.908	0.905	0.900	0.911
g4			0.996	0.842	0.840	0.835	0.834
g5				0.993	0.992	0.986	0.979
g6					0.845	0.838	0.831
g7						0.899	0.897
g8							0.924

Table 3e. Estimates of reliability and SE after each item

		SUM1-2	SUM1-3	SUM1-4	SUM1-5	SUM1-6	SUM1-7	SUM1-8
Theta	$\rho_{TH}=\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}\lambda_{i\theta}^2}\right)$	0,098	0,264	0,326	0,319	0,344	0,353	0,444
Omega	$\rho_{\omega}=\frac{\left(\sum_{i=1}^{k}\lambda_{i\theta}\right)^2}{\left(\sum_{i=1}^{k}\lambda_{i\theta}\right)^2+\sum_{g=1}^{k}\left(1-\lambda_{i\theta}^2\right)}$	0,097	0,293	0,326	0,312	0,336	0,343	0,422
ThetaG	$\rho_{THG}=\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}G_{iX}^2}\right)$	1	0,998	0,993	0,938	0,945	0,952	0,961
OmegaG	$\rho_{\omega G}=\frac{\left(\sum_{i=1}^{k}G_{iX}\right)^2}{\left(\sum_{i=1}^{k}G_{iX}\right)^2+\sum_{g=1}^{k}\left(1-G_{iX}^2\right)}$	1	0,998	0,995	0,952	0,956	0,961	0,967
SE (Theta)	$\sigma_X\sqrt{1-\rho_{TH}}$	0,048	0,050	0,119	0,395	0,402	0,441	0,425
SE (Omega)	$\sigma_X\sqrt{1-\rho_{\omega}}$	0,048	0,048	0,119	0,400	0,406	0,448	0,442
SE (ThetaG)	$\sigma_X\sqrt{1-\rho_{THG}}$	0	0,000	0,001	0,036	0,033	0,032	0,030
SE (OmegaG)	$\sigma_X\sqrt{1-\rho_{\omega G}}$	0	0,000	0,001	0,028	0,027	0,027	0,025

Figure 10. Estimates of reliability and standard error at each step of adding items in a test



First note to make is that the traditional estimators theta and omega cannot detect the fact in the dataset that the lowest scoring test-takers are systematically scoring lower also in the items. Hence, the low reliability (0.097−0.423). From this viewpoint, estimates by *G* are closer the truth: on average, around

94–100% of the test-takers are logically ordered in all items in each step of adding the items to the test.[3] This is seen also in the high magnitude of the estimates of reliability using *G* as the weight factor (0.961–1.000). Notably, the estimates would be slightly lower although at the same range if *RPC* or *D* would be used in DCERs instead of *G*; with all 8 items, $\rho_{TH_{RPCiX}}$=0.869, $\rho_{TH_{DiX}}$=0.937, $\rho_{\omega_{RPCiX}}$=0.895, and $\rho_{\omega_{DiX}}$=0.947 (see Metsämuuronen, 2022b). Second, because of the notable deflation in the estimates of reliability by the traditional estimators, the standard errors based on their estimates are notably inflated (0.506–0.524 after fifth item) in comparison with those based on DCERs (0.029–0.036). Third, in the dataset, the standard errors are rather stable after the fifth item is added to the test. The estimates based on DCERs are notably more stable than those based on the traditional estimators. Obviously, systematic studies of the phenomena discussed here are beneficial. The empirical results give a hint though that using DCERs in estimating the conditional standard errors is worth studying more.

6.3 Practical implications related to the results

To outline the practical suggestions based on the literature and simulation, if one is willing to maximize the probability that the estimate of reliability would be as close as possible the true, population value, it is recommended to

(1) select a proper measurement model fitting the dataset,

(2) select the best option of the estimators within the model selected (although this may not affect much),

(3) consider using weighted score instead of the raw score (although this may not affect much),

(4) ponder whether items with polytomous or continuous scales could be used instead of binary ones (although this may not affect much),

(5) consider raising the sample size higher than 25–50 test-takers (this may have a remarkable effect),

(6) consider constructing the score so that it would have 15 categories or more (this may have a remarkable effect),

(7) consider changing the weight factor in the traditional estimators of reliability to a one with less mechanical error (this may have a remarkable effect), and

(8) use items with extreme difficulty level in the test to give test-takers possibility to show at least some achievement (easy items) or how far they can reach (difficult items). However, when items with extreme difficulty levels are used, consider using estimators from the extended family of DCERs instead of the traditional estimators to estimate the reliability (this may have a remarkable effect).

These are well-known facts within the professionals working with testing. Anyhow, following these basic principles maximizes the probability to obtain as accurate estimate of reliability as possible for varying purposes of reliability.

6.4 Known limitations

An obvious limitation of the study is that a simulation with real-world items has its own limitations. Although the numbers of subtests (n = 1,440) and items (k = 14,882) used in the study are rather convincing, those are based on *one* basic dataset. Results may have been somewhat different if truly polytomous test items were used in the simulation. Replications of the design or another approach with a more independent estimates may increase our knowledge of the relation between the estimators.

The analysis did not concern very difficult and very easy tests; using the original dataset, this would have required very short tests with binary items. The results here and the empirical results by e.g. Metsämuuronen and Ukkola (2019) give a hint that the behaviour of the estimators with very difficult and easy tests would show remarkable deflation in estimates. Also, the simulation included only tests with minimum 20 points as the maximum score; the behaviour of the estimators

[3] Because of the relation between *G* and Jonckheere–Terpstra test statistic (see Metsämuuronen, 2021b), probability that the test-takers are in an ascending order in an item after they are ordered by the score is $p = 0.50 \times G + 0.50$. Here, the average of *G*s is used in calculation.

may be different with very short tests. Some ideas of the radical deflation in the estimates of reliability by the traditional estimators with tests of extreme difficulty level were given in the discussion section. Studies in this respect would be beneficial.

The comparison of the DCERs and traditional theta, omega, and rho was done by using the raw score as the manifestation of θ instead of the traditional factor score. If used the factor score variables as the manifestation of θ, there would be a better comparison between the estimators. Studies in this respect may increase our knowledge of the matter (see some comparison in Metsämuuronen, 2022b though). Also, comparisons of estimates by ordinal alpha and theta by Zumbo and colleagues (2007) and DCERs discussed in this article would, be beneficial.

References

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika 18*, 1–14. https://doi.org/10.1007/BF02289023

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. (1968). Alpha-maximized factor analysis (Alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika, 33*(3), 335–345. https://doi.org/10.1007/BF02289328.

Bridgeman, B. (2016). Can a two-question test be reliable and valid for predicting academic outcomes? *Educational Measurement: Issues and Practice, 35*(4), 21–24. https://doi.org/10.1111/emip.12130

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

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

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

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

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

Davenport, E. C., Davison, M., L., Liou, P.-Y., & Love, Q., U. (2016). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice 34*(4), 4–9. https://doi.org/10.1111/emip.12095

Davenport, E. C., Davison, M., L., Liou, P.-Y., & Love, Q., U. (2016). Easier said than done: Rejoinder on Sijtsma and on Green and Yang. *Educational Measurement: Issues and Practice, 35*(1), 6–10. https://doi.org/10.1111/emip.12106

Feldt, L. S. (1975) Estimation of reliability of a test divided into two parts of unequal length. *Psychometrika 40*, 557–561. https://doi.org/10.1007/BF02291556

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (ed.) *Educational Measurement*. 3rd edition. American council of education. Series of Higher Education. Oryx Press.

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. https://timssandpirls.bc.edu/timss2019/methods/chapter-14.html

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

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika, 48*, 99–111. https://doi.org/10.1007/BF02314679

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. http://dx.doi.org/10.1007/s11336-008-9098-4

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice, 34*(4), 14–20. http://dx.doi.org/10.1111/emip.12100

Greene, V. L., & Carmines, E. G. (1980). Assessing the Reliability of Linear Composites. *Sociological Methodology, 11*,160–17. https://doi.org/10.2307/270862

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

Guttman, L. (1941). The qualifications of a class of attributes: a theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment*. Social Science Research Council, Bulletin 48, 321–345.

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). Lincoln, IL: Scientific Software International, Inc.

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

Horst, P. (1951). Estimating the total test reliability from parts of unequal length. *Educational and Psychological Measurement, 11*(3), 368–371. https://doi.org/10.1177/001316445101100306

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, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review, 11*, 179–188. https://doi.org/10.1007/s12564-009-9062-8

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

Livingston, S. A., & Dorans, N. J. (2004). *A graphical approach to item analysis*. (Research Report No. RR-04-10). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01937.x

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.

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

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

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th Edition). Harcourt Brace College Publishers.

Metsämuuronen, J. (2016). Item–total correlation as the cause for the underestimation of the alpha estimate for the reliability of the scale. *GJRA - Global Journal for Research Analysis*, *5*(1), 471–477. https://www.worldwidejournals.com/global-journal-for-research-analysis-GJRA/file.php?val=November_2016_1478701072__159.pdf

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

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 Methodology, 6*(1), 207–221. https://doi.org/10.12973/ijem.6.1.207

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. (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). 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, 91–130. https://doi.org/10.1007/s41237-022-00158-y,

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

Metsämuuronen, J. (2022c). Attenuation-corrected reliability and some other MEC-corrected estimators of reliability. *Applied Psychological Measurement*. (In printing)

Metsämuuronen, J. (2022d). 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://link.springer.com/content/pdf/10.1007/s41237-022-00162-2.pdf

Metsämuuronen, J. & Ukkola, A. (2019). *Methodological solutions of zero level assessment* (*Alkumittauksen menetelmällisiä ratkaisuja*). Publications 18:2019. Finnish Education Evaluation Centre. [in Finnish] 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*, 1–13. https://doi.org/10.1007/BF02289400

Olsson, 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. (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. (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

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42*, 549–565. https://doi.org/10.1007/BF02295978

Raykov, T. (1997a). Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence for Fixed Congeneric Components. *Multivariate Behavioral Research*, *32*(4), 329–354. http://doi.org/10.1207/s15327906mbr3204_2.

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

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*, 57–74. http://dx.doi.org/10.1207/s15327906mbr1401_4

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

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

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

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

Smith, J. K. (2005). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice, 22*(4), 26–33. https://doi.org/10.1111/j.1745-3992.2003.tb00141.x

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. (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

Stouffer, S. A. (Ed.) (1950). Measurement and prediction. *Studies in social psychology in World war II*, Vol IV. Princeton, N.J.: Princeton university press.

Ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*(4), 575–579. http://dx.doi.org/10.1007/BF02293815.

Thompson, G. H. (1940). Weighting for battery reliability and prediction. *British Journal of Mathematical and Statistical Psychology,30*/4), 357–360. https://doi.org/10.1111/j.2044-8295.1940.tb00968.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

Warrens, M. J. (2016). A comparison of reliability coefficients for psychometric tests that consist of two parts. *Advances in Data Analysis and Classification*, *10*, 71–84. https://doi.org/10.1007/s11634-015-0198-6

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*, 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. http://dx.doi.org/10.4135/9781412961288.n309

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's *a*, Revelle's *β*, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133. http://dx.doi.org/10.1007/s11336-003-0974-7

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. (2022). How to obtain the most error-free estimate of reliability? Eight sources of deflation in the estimates of reliability. *Practical Assessment, Research & Evaluation*, 27(10). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/10/

Corresponding Author:

Jari Metsämuuronen

1) Finnish Education Evaluation Centre (FINEEC)
P.O. Box 380 (Hakaniemenranta 6)
FI-00531 HELSINKI

2) Centre for Learning Analytics,
FI-20014, University of Turku, Finland

Email: jari.metsamuuronen [at] gmail.com

Appendix A. List of abbreviations used in the article

General abbreviations	
df(g)	degrees of freedom of item = number of categories–1
df(X)	degrees of freedom of the score = number of categories–1
k	number of items
p	probability, proportion of correct answers
MEC	mechanical error in estimates of correlation
MLE	maximum likelihood estimation
PAF	principal axis factoring
SE	Standard Error
S.E.m	Standard error of measurement
REL	reliability
Concepts related to variables	
X	observed score variable
g	observed item
T	unobserved true score
E	unobserved error score
x_i	observed value of X
t_i	observed value of T
e_i	observed value of E
Types of score variables	
θ	latent variable
θ_{RAW}, θ_X, X	latent variable manifested as a raw score
θ_{PC}	latent variable manifested as a principal component score
θ_{FA}	latent variable manifested as a factor score variable
θ_{IRT}	latent variable manifested as a theta score formed by the item response theory (IRT) or Rasch modelling
θ_{NonL}	latent variable manifested as a nonlinear compilation of items
Estimators of reliability	
ϱ_{BS}	Brown–Spearman prediction formula
ϱ_{FR}	Flanagan–Rulon prediction formula
ϱ_{GLB}	greatest lower bound reliability
ϱ_{LLB}	lowest lower bound of reliability
ϱ_{KR20}	Kuder and Richardson formula 20
ϱ_{KR21}	Kuder and Richardson formula 21
λ_1—λ_6	Guttman family of estimators
ϱ_{α}	coefficient alpha, Cronbach alpha
ϱ_{TH}	coefficient theta, Armor theta
ϱ_{ω}	coefficient omega, McDonald omega total
ϱ_{MAX}	coefficient rho, maximal reliability, Raykov rho, Hancock H
ϱ_{H}	Horst coefficient
ϱ_{AF}	Angoff–Feldt coefficient
ϱ_{β}	Raju coefficient, Raju's β

ϱ_{GF}	Gilmer–Feldt coefficient
Estimators of correlation	
PMC = ρ_{XY}	product-moment correlation coefficient between variables *X* and *Y*, Pearson correlation
$Rit = \rho_{iX}$	item–total correlation, item–score correlation, a special case of PMC
λ_i	factor loading, principal component loading
R_{PC}	polychoric correlation
R_{REG}	bi- and polyreg coefficient
G	Goodman–Kruskal gamma
G_2	dimension-corrected *G*
D	Somers delta
D_2	dimension-corrected *D*
R_{AC}	attenuation-corrected PMC
E_{AC}	attenuation-corrected eta
Concepts related to deflation-corrected estimators of reliability	
DCER	Deflation-corrected estimator of reliability
w_i, $w_{i\theta}$	weight factor, correlation between an item *i* and the latent variable manifested as a score variable θ
$\rho_{\alpha\_wi\theta}$	deflation-corrected alpha
$\rho_{TH\_wi\theta}$	deflation-corrected theta
$\rho_{\omega\_wi\theta}$	deflation-corrected omega
$\rho_{MAX\_wi\theta}$	deflation-corrected rho