Behaviormetrika (2022) 49:91–130
https://doi.org/10.1007/s41237-022-00158-y

ORIGINAL PAPER

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

Jari Metsämuuronen[1,2]

Received: 6 July 2021 / Accepted: 11 January 2022 / Published online: 10 February 2022

Abstract
Estimates of reliability by traditional estimators are deflated, because the item-total or item-score correlation (*Rit*) or principal component or factor loading (λ_i) embedded in the estimators are seriously deflated. Different optional estimators of correlation that can replace *Rit* and λ_i are compared in this article. Simulations show that estimators such as polychoric correlation (R_{PC}), gamma (G), dimension-corrected G (G_2), and attenuation-corrected *Rit* (R_{AC}) and *eta* (E_{AC}) reflect the true correlation without any loss of information with several sources of technical or mechanical error in the estimators of correlation (MEC) including extreme item difficulty and item variance, small number of categories in the item and in the score, and the varying distributions of the latent variable. To obtain deflation-corrected reliability, R_{PC}, G, G_2, R_{AC}, and E_{AC} are likely to be the best options closely followed by r-bireg or r-polyreg coefficient (R_{REG}).

Keywords Reliability · Item-total correlation · Polychoric correlation · Goodman–Kruskal gamma · Somers delta · Kendall *tau-b* · r-Polyreg coefficient · Attenuation-corrected correlation

1 Introduction: deflation in reliability and correlation as phenomena

One of the most enduring areas of interest related to measurement modelling is the underestimation of reliability of the test score. Guttman (1945) was the first to show that the estimate of reliability obtained by the formula known today as coefficient

Communicated by Kohei Adachi.

✉ Jari Metsämuuronen
jari.metsamuuronen@gmail.com

[1] Finnish Education Evaluation Centre, P.O. Box 380, (Hakaniemenranta 6), FI-00531 Helsinki, Finland

[2] Centre for Learning Analytics, University of Turku, Turku, Finland

alpha (α; chronologically, Kuder and Richardson 1937; Jackson and Ferguson 1941; Guttman 1945) or Cronbach's alpha (Cronbach 1951) is always lower than the population reliability. This observation is traditionally referred to as "lower bound of reliability" (e.g., Gulliksen 1950; Guttman 1945) or "reliability in the case of (essential) tau–equivalence situation" (e.g., Novick and Lewis 1967). Under-estimation in the estimates by α has been connected to a simplified assumption of the classical test theory including violations in tau–equivalence and latent normality, unidimensionality, and uncorrelated errors (e.g., Green and Yang 2009, 2015; McNeish 2017; Trizano-Hermosilla and Alvarado 2016).

Usually, this *attenuation* in reliability is seen as a natural consequence of random errors in the measurement. However, a less discussed challenge in the estimates by the traditional estimators of reliability is that their estimates may be radically *deflated* caused by artificial systematic errors during the estimation or (see the discussion of the terms in, e.g., Chan 2008; Gadermann et al. 2012; Lavrakas 2008). Empirical examples (see Sect. 1.2) show that, in very easy and very difficult tests and tests with incremental difficulty level including both easy and difficult items, the estimates of reliability may be deflated by 0.40–0.60 units of reliability (see, e.g., Gadermann et al. 2012; Metsämuuronen and Ukkola 2019; Zumbo et al. 2007; see Sect. 1.2). This kind of deflation is caused by a phenomenon called the technical or mechanical error in estimates of correlation (MEC) discussed, specifically by Metsämuuronen (e.g., 2022a, b, c). In measurement modelling settings between the test items (g_i) and the latent trait θ manifested as a score variable (X), MEC refers to such technical reasons in estimators of correlation as number of categories or item difficulty to underestimate the true correlation between g_i and X. These kinds of technical reasons cause (mechanical) attenuation in correlation in general and, specifically, in product–moment correlation coefficient (PMC; Pearson 1896 onwards) embedded in the most used estimators of reliability either in the form of item-total correlation or, more generally, item-score correlation (*Rit*) or principal component- or factor loading (λ_i). It is known that PMC always underestimates the true correlation in an obvious manner when the number of categories of the variables of interest is not equal (see algebraic reasons in, e.g., Metsämuuronen 2016, 2017, 2020c; and simulations in Martin 1973, 1978; Olsson 1980; Metsämuuronen 2021a), and this is always the case with item and score. This phenomenon and its consequences in the estimates of reliability are discussed briefly below.

1.1 Sources of MEC and attenuation in estimators of correlation

Attenuation in correlation can be partly explained by the phenomenon called range restriction or restriction in range (see the literature in, e.g., Mendoza and Mumford 1987; Sackett and Yang 2000; Sackett et al. 2007; Schmidt et al. 2008; see also Meade 2010; Walk and Rupp 2010). Restriction in range refers to a phenomenon when only a portion of the range of values of the (latent) variable is actualized in the sample as is the case, for example, when a highly selected sample participates in an entrance test causing the sample variance to reduce in comparison with the population variance. This leads to inaccuracy in the estimates related to the score.

Nevertheless, even if no restriction in range is obtained per se, PMC always underestimates the true correlation in an obvious manner when the number of categories of the variables of interest is not equal as is always the case with item as discussed above. Several other reasons for the attenuation can be detected and 11 sources of MEC are discussed in what follows.

Generalizing from the simulations of multiple sources of MEC by Metsämuuronen (2020b, 2021a; see also the empirical section below), the technical attenuation in *Rit* arises, at least, from six sources. First, *Rit* tends to underestimate the true correlation always when the number of categories in the variables differs from each other; that is, (1) *Rit* is *sensitive to discrepancy in scales*. This source of MEC may be related to RR, and this always happens between an item and a score. Second, *Rit* tends to underestimate the true correlation the more extreme is the item difficulty leading to reduced item variance; that is, (2) *Rit* is *sensitive to item difficulty and item variance*. This causes drastic underestimation with very easy and very difficult item; the loss on information approximates 100% depending on the sample size. Third, *Rit* tends to underestimate the true correlation more when the distribution of the latent variable is normal or skewed than when it is even; that is, (3) *Rit* is *sensitive to the distribution of the latent variable*. Fourth, *Rit* tends to underestimate the true correlation more the less categories there are in the item; that is, (4) *Rit* is *sensitive to the number of categories in the item*. Fifth, *Rit* tends to underestimate the true correlation more the less categories there are in the score; that is, (5) *Rit* is *sensitive to the number of categories in the score*. Sixth, *Rit* tends to underestimate the true correlation more the less there are items forming the score, because this has a strict connection to the number of categories in the scale of the score; that is, (6) *Rit* is *sensitive to the number of items forming the score*. Seventh, bound to sources 5 and 6, *Rit* tends to underestimate the true correlation more the more there are tied cases in the score; that is, (7) *Rit* is *sensitive to the number of tied cases in the score*. In the empirical section, these sources of MEC are examined in Study 1 using a theoretical dataset.

The sources of MEC above are not the only possible ones, although they strictly affect *Rit*. Generalizing from Metsämuuronen (2020b, 2021a), sources of MEC also include (8) *symmetric nature* of the coefficient, and (9) latent *linear nature* of the coefficient. The latter source affects, specifically, such estimators based on probability as Kendall *tau-b* (Kendall 1948), Goodman–Kruskal gamma (*G*; Goodman and Kruskal 1954), and Somers delta (*D*; Somers 1962). The effect of these sources is examined in Study 2 by using a real-world dataset.

Two specific types of characteristics related to an estimator as a suitable option for *Rit* are (10) *instability of the estimator to reflect the population correlation*, and (11) *the tendency to overestimate the population correlation*. These are examined in Study 3 by using the same real-world dataset as in Study 2.

To illustrate the differences in magnitudes of error (e) caused by the mechanical error (MEC) by different estimators ($w_{i\theta}$) of item–score correlation ($e_{wi\theta\_\mathrm{MEC}}$), let us consider Fig. 1 based on a dataset used in Study 1 where a pair of identical, normally distributed variables with (obvious) perfect correlation is manipulated, so that one variable is dichotomized (item g) and the other is polytomized to 21 categories (score X). As an example, the outcome of a binary item with the proportion of 1 s

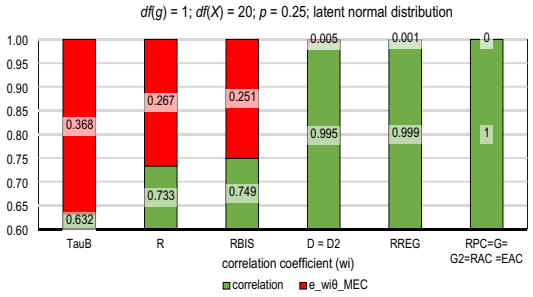

Fig. 1 Magnitude of MEC in different estimators

TauB = Kendall's tau-b; R = Rit = PMC; RBIS = biserial correlation; D = Somers delta (X dependent); D2 = dimension-corrected D; RREG = bi-reg correlation; RPC = polychoric correlation; G = Goodman-Kruskal gamma; G2 = dimension-corrected G; RAC = attenuation-corrected Rit; EAC = attenuation-corrected eta

being $p=0.25$ (or $p=0.75$) is seen in Fig. 1. The estimators are discussed later with the literature.

We note that, of the estimators in comparison, Kendall's *tau-b* and *Rit* cannot reach the (obvious) perfect correlation between the binary and polytomized version of the same variable and, hence, the magnitude of error related to MEC is the highest ($e_{Tau\text{-}bi\theta\_MEC}=0.37$ and $e_{Ri\theta\_MEC}=0.27$ units of correlation), while R_{PC}, G, G_2, R_{AC}, and E_{AC} reach the perfect correlation. In the empirical section, selected estimators of correlation are compared to see to what extent they are affected by the 11 sources of MEC discussed above.

1.2 Practical consequences of MEC in the estimators or reliability

The deflation in estimators of correlation caused by MEC has led to discussion about deflation-corrected estimators of reliability (DCER). These are divided into MEC-corrected estimators of reliability (MCER; Metsämuuronen 2021a, 2022a) where the traditional estimator of correlation (PMC) is replaced by totally different estimator (e.g., R_{PC}, G, or D) and attenuation-corrected estimators of reliability (ACER; Metsämuuronen 2022b, c) where a relevant attenuation-corrected estimator of correlations (e.g., R_{AC} or E_{AC}) is used instead of the traditional estimator. The discussion is summarized here to motivate a comparison of suitable alternatives of estimates for *Rit* and λ_i in the estimators of reliability.

Empirical results indicate that MEC in *Rit* may have a radical effect in reliability. Gadermann et al. (2012), Metsämuuronen (2022a, b), and Metsämuuronen and Ukkola 2019 report that the traditional coefficient alpha and maximal reliability may underestimate reliability by 0.40–0.60 units of reliability. The reduction is notable and worth studying. The main reason for the deflation in estimates by the widely used traditional estimators of reliability is the poor behavior of *Rit* with items of extreme difficulty level (see, simulations by, e.g., Metsämuuronen 2020a, 2021a; see also the empirical section below). Attenuation in reliability is caused by the fact that, on the one hand, *Rit* is *visible* in such classic estimators of reliability as Kuder and Richardson formulae 20 and 21 (Kuder and Richardson 1937) and coefficient alpha. Common to these classic estimators is that the variance of the test score (σ_X^2) inherited from the basic definition of reliability

($REL = \sigma_T^2/\sigma_X^2 = 1 - \sigma_E^2/\sigma_X^2$) is visible in the formula, and σ_X^2 can be expressed by item variances σ_i^2 and PMC = $Rit = \rho_{iX}$: $\sigma_X^2 = \left(\sum_{i=1}^k \sigma_i \times \rho_{iX}\right)^2$ (Lord and Novick 1968), where k refers to the number items in the compilation. Then, the coefficient alpha can be expressed as

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

(Lord and Novick 1968) where PMC is visible. On the other hand, PMC is *embedded* in the estimators based on principal component- and factor analysis, because the principal component and factor loadings λ_i are (essentially) correlations between item and the score variable (see Cramer and Howitt 2004; Kim and Mueller 1978; Yang 2010). This concerns such estimators as Armor's theta (ρ_{TH}; Armor 1973; see also Kaiser and 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 and which maximizes alpha (Greene and Carmines 1980) as well as McDonald's omega total ($\rho_{\omega_{Total}} = \rho_\omega$; Heise and Bohrnstedt 1970; McDonald 1970)

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

and maximal reliability (ρ_{MAX}; e.g., Raykov 2004 onwards), based on the conceptualization suggested by Li et al. (1996) and Li (1997)

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

(e.g., Cheng et al. 2012) where λ_i are factor loadings.

Estimators (2), (3), and (4) are based on a (simplified, one-latent factor) measurement model where the observed responses in g_i (x_i) are explained by a latent variable θ with a linking element λ_i between θ and g_i where $-1 \le \lambda_i \le +1$, and error related to the model (e_i)

$$x_i = \lambda_i \theta + e_i \tag{5}$$

(e.g., Cheng et al. 2012; McDonalds 1985, 1999). Traditionally, the model in Eq. (5) assumes that λ_i is MEC-free. This is, however, a too optimistic assumption, because loading is, essentially, item–score correlation as discussed above, and the deflation in reliability may be substantial.

1.3 Conceptual and theoretical consequences of MEC in the estimators or reliability

We keep in mind (Fig. 1) that the magnitude of error caused by MEC varies coefficient-wise (w) and item-wise (i) and score variable-wise (θ). To formalize the error element $e_{wi\theta\ \_\mathrm{MEC}}$ related to MEC to the measurement model, let us reconceptualize Eq. (5) as a more general model

$$x_i = w_{i\theta}\theta + e_i, \tag{6}$$

where θ may be a relevantly formed compilation of items such as raw score (θ_{RAW}), principal component score (θ_{PC}), factor score (θ_{FA}), score formed by item response theory (IRT) or Rasch modelling (θ_{IRT}), or a nonlinear compilation of varied kind (θ_{NonL}). The weight element w_i need not to be bound exclusively to the mechanics of principal component- or factor analysis. However, it makes sense that w_i is (essentially) a coefficient of correlation ($-1 \le w_i \le +1$) such as *Rit*, *G*, *D*, *tau-b*, or polychoric correlation (R_{PC}; Pearson 1900, 1913) discussed later in this article, or the traditional factor or principal component loadings (λ_i).

If $w_{i\theta}$ includes MEC, as it typically does when using the traditional estimators of reliability,[1] the observed estimate by the MEC-defected (MECD) estimator of correlation ($w_{i\theta\_MECD}$) such as *tau-b* or *Rit* underestimates the true, MEC-free (MECF) correlation ($w_{i\theta\_MECF}$), that is

$$w_{i\theta\_MECD} = w_{i\theta\_MECF} - e_{wi\theta\_MEC} \tag{7a}$$

or

$$w_{i\theta\_MECF} = w_{i\theta\_MECD} + e_{wi\theta\_MEC}, \tag{7b}$$

where the error related to MEC is positive ($e_{wi\theta\_MEC} > 0$). Equation (7b) suggests to reconceptualize the classic relation of the observed score (X), true score (T), and error (E), that is, $X = T + E$ (Gulliksen 1950) into a form

$$X = T + \left(E_{Random} + E_{MEC}\right) \tag{8}$$

and to rewrite the measurement model in Eq. (6) as

$$x_i = \left(w_{i\theta\_MECD}\right) \times \theta + \left(e_{i\_Random} + e_{wi\theta\_MEC}\right). \tag{9}$$

[1] Recall that some traditional estimators of correlation used as the linking element in measurement modelling settings such as biserial (R_{BS}) and polyserial correlation (R_{PS}) coefficients (Pearson, 1909) tend to give obvious *over*estimates to the extent of out-of-range values (R_{BS}, $R_{PS} < +1.252$) if we use the traditional way in estimation (see, Drasgow, 1986) and if PMC and the item variance are high to start with (e.g., Clemans, 1958; see also Metsämuuronen, 2020a). Some researchers have argued that *G* also overestimates correlation (e.g., Higham and Higham, 2019; Kvålseth, 2017). However, there does not seem to be an "inflation" per se in *G* (see Gonzalez and Nelson, 1996; Metsämuuronen, 2021b); the higher values of *G* in comparison with *D* and *tau-b* are caused by its hidden directional nature and by a different way of thinking about the probability, the same logic of probability as used in the traditional sign test and Wilcoxon signed-rank test (see Metsämuuronen 2021a,b).

It may be too optimistic to claim that some estimator of correlation would be totally MEC-*free*. Hence, it may better to lose the requirement from MEC-free conditions to MEC-*corrected* (MECC) conditions where $e_{wi\theta\_MEC} \approx 0$. If we use options of weight coefficient which would lead us to the condition of $e_{wi\theta\_MEC} \approx 0$, because of Eqs. (7b) and (9), this would lead us to a model where the estimate by the selected weight factor would be as near the MEC-free condition as possible, that is

$$\begin{aligned} x_i &= w_{i\theta\_MECC} \times \theta + \left(e_{i\_Random} + e_{wi\theta\_MEC}\right) \\ &\approx w_{i\theta\_MECF} \times \theta + e_{i\_Random}. \end{aligned} \tag{10}$$

Equation (10) strictly leads us to the traditional measurement model of summed items (see e.g., Cheng et al. 2012). Knowing that all generally used estimators of correlation give identical estimate of the correlation for original variables (g_i and θ) and for the standardized versions of the variables ($STD(g_i)$ and $STD(\theta)$), without loss of generality, we can assume that, from the viewpoint of measurement modelling, g_i and θ are standardized, $x_i, \theta \sim N(0, 1)$. Then, assuming that item-wise random errors do not depend on the true scores, the item-wise and score-wise MEC-corrected error variance ($\psi^2_{i\theta\_MECC}$) is

$$\psi^2_{i\theta\_MECC} = VAR(e_i) = VAR(x_i) - \left(w_{i\theta\_MECC}\right)^2 \times VAR(\theta) = 1 - w^2_{i\theta\_MECC}, \tag{11}$$

that is, $e_{wi\theta\_MECC} \sim N\left(0, \psi^2_{i\theta\_MECC}\right)$ where $\psi^2_{i\theta\_MECC} = 1 - w^2_{i\theta\_MECC}$. Then, the MEC-corrected relation $X = T + E_{Random} + E_{MEC} - E_{MEC} = T + E_{Random}$ concerning the score variable can be rewritten as

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} w_{i\theta\_MECC} \times \theta + \sum_{i=1}^{k} e_{i\_Random}, \tag{12}$$

where k is the number of items in the compilation. Consequently, the MEC-corrected error variance of the test score can be written as

$$\sum_{i=1}^{k} \psi^2_{i\theta\_MECC} = \sum_{i=1}^{k} \left(1 - w^2_{i\theta\_MECC}\right) \tag{13}$$

instead of the traditional MEC-defected error variance

$$\sum_{i=1}^{k} \psi_i^2 = \sum_{i=1}^{k} \left(1 - \lambda_i^2\right) \tag{14}$$

used in the traditional estimators of omega and rho in Eqs. (3) and (4).

Replacing the MEC-defected estimators of correlation *Rit* and λ_i in the traditional estimators of reliability with MEC-corrected estimators of correlation leads us to such (theoretical) families of deflation-corrected estimators of reliability (DCER) as MEC-corrected alpha

$$\rho_{\alpha\_MECC} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times w_{i\theta\_MECC}\right)^2}\right) = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times w_{i\theta}\right)^2}\right), \tag{15}$$

MEC-corrected theta

$$\rho_{TH\_MECC} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} w_{i\theta\_MECC}^2}\right) = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} w_{i\theta}^2}\right), \tag{16}$$

MEC-corrected omega total

$$\rho_{\omega\_MECC} = \frac{\left(\sum_{i=1}^{k} w_{i\theta\_MECC}\right)^2}{\left(\sum_{i=1}^{k} w_{i\theta\_MECC}\right)^2 + \sum_{i=1}^{k}\left(1 - w_{i\theta\_MECC}^2\right)} = \frac{\left(\sum_{i=1}^{k} w_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} w_{i\theta}\right)^2 + \sum_{i=1}^{k}\left(1 - w_{i\theta}^2\right)}, \tag{17}$$

and MEC-corrected maximal reliability

$$\rho_{MAX\_MECC} = \frac{\sum_{i=1}^{k}\left(w_{i\theta\_MECC}^2 \Big/ \left(1 - w_{i\theta\_MECC}^2\right)\right)}{1 + \sum_{i=1}^{k}\left(w_{i\theta\_MECC}^2 \Big/ \left(1 - w_{i\theta\_MECC}^2\right)\right)} = \frac{\sum_{i=1}^{k}\left(w_{i\theta}^2 \big/ \left(1 - w_{i\theta}^2\right)\right)}{1 + \sum_{i=1}^{k}\left(w_{i\theta}^2 \big/ \left(1 - w_{i\theta}^2\right)\right)}. \tag{18}$$

The task is to find w_i where the quantity of MEC is as small as possible to approach as close as possible the MEC-free estimates of reliability. Some practical options are suggested after the empirical section.

1.4 Research question

Although rigorous studies have been done on PMC and selected alternatives (e.g., Anselmi et al. 2019; Martin 1973, 1978; Olsson 1980; Metsämuuronen 2020a, b, 2021a), from the viewpoint of conjoint elements of MEC, these tend to be fragmentary. Systematic studies of the several simultaneous elements of MEC would enrich our knowledge of the phenomenon. This article intends to partially cover for this lack of knowledge. The effect of 11 sources of MEC on eight benchmarking alternative estimators of correlation is studied in three sub-studies. The purpose is to quantify the effect of MEC in selected estimators of correlation and to select the most potential options to be used as less MEC-affected weight factors w_i instead of *Rit* in MEC- and attenuated-corrected estimators of reliability. Focused research questions are presented with the separate studies.

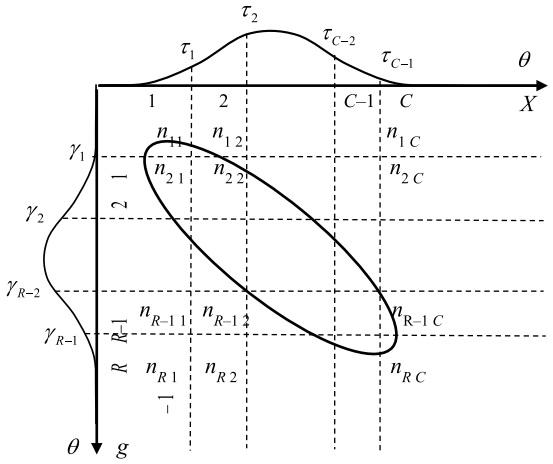

Fig. 2 A latent variable θ manifested in two different ordinal scales

2 General methodological issues

2.1 Statistical model related to the estimators of correlation in measurement modelling settings

Assume that the observed values in item g (x_i) and score variable X (y_i) with $r=1$, …, R and $c=1$, …, C distinctive ordinal or interval categories, respectively, share the common latent variable θ and, usually, $R<<C$. The threshold values of θ for each category in g and X are denoted by γ_i and τ_j, respectively. Then, g and X are related to θ, so that $g=x_i$, if $\gamma_{i-1}\leq\theta<\gamma_i$, $i=1, 2,\ldots, R$ and $X=y_j$, if $\tau_{j-1}\leq\theta<\tau_j$, $j=1, 2, \ldots, C$, and $\gamma_0 = \tau_0 = -\infty$ and $\gamma_R = \tau_C = +\infty$. The statistical model is illustrated in Fig. 2 imitating, unconventionally, the logic of a two-way contingency table.

2.2 Estimators of correlation in comparison

Options for the coefficients to be used as weight factor w_i are many—usually, these have been discussed under the topic of "item discrimination power" (see, e.g., Oosterhof 1976, who compared 19 of these). Some estimators based on the *covariation* with latent *trigonometric nature* are already named above such as *Rit*, and R_{PC}. This group may also include a coefficient called an *r*-biserial and *r*-polyreg correlation (R_{REG}) based on the regression coefficients (see Livingston and Dorans 2004; Moses 2017). Different types of nonparametric estimators are based on the *probability* with latent *linear nature* including *tau-b*, *G,* and *D* discussed above. In-between the trigonometric and linear nature falls a pair of estimators called dimension-corrected *G* (G_2; Metsämuuronen 2021a) and dimension-corrected *D* (D_2; Metsämuuronen 2020b, corrected in 2021a) where the linear nature of *G* and *D* is transformed into a more trigonometric one—Metsämuuronen (2021b) calls this *semi-trigonometric nature*. One pair of new estimators with unknown latent nature are

attenuation-corrected PMC (R_{AC}) and attenuation-corrected *eta* (E_{AC}) suggested by Metsämuuronen (2022b, c). These estimators are briefly discussed in what follows.

2.2.1 Product–moment correlation coefficient

PMC is used for estimating the correlation between two observed variables. In measurement modelling settings with item (g) and score (X), it can be expressed as

$$PMC = Rit = \rho_{gX} = \frac{\sigma_{gX}}{\sigma_g \sigma_X}, \tag{19}$$

where σ_g and σ_X refer to the standard deviations and σ_{gX} is the item–score covariance.

When it comes to the directional or symmetric nature of *Rit*, Metsämuuronen (2020c, see also 2022c) showed that PMC has a hidden directional nature, so that the variable with a wider scale (X) explains the response pattern in the variable with a narrower scale (g). This direction makes sense in measurement modelling settings where we assume that the latent trait θ manifested as the score variable (X) explains the response pattern in the item (g) and not opposite, that is, the direction of "g given X" from the conditions viewpoint. This is same direction familiar from the traditional use of *eta* squared usually labelled as "score dependent" in settings related to general linear modelling (GLM; see the discussion of naming the directions in, e.g., Metsämuuronen 2020a, 2022c). Because its characteristics are well studied, *Rit* is used as a benchmarking estimator: if the magnitude of an estimate by some estimator is *lower* than that by *Rit*, this indicates *obvious underestimation*.

2.2.2 Polychoric correlation

R_{PC} is used for estimating the inferred correlation between two unobservable latent variables that are truncated to ordinal (or interval)-scaled forms. R_{PC} differs from PMC in that there is no closed-form expression for the relation between g and X. Instead, several alternatives for the estimation process are suggested (see some options in Drasgow 1986; Olsson et al. 1982) and these may produce slightly different estimates. One of these options is the two-step estimator by Martinson and Hamdan (1972), which is used in this article. In its simplified form (see, Zaionts 2021), the task is to find the value of PMC that maximizes the log-likelihood function LL where

$$LL = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij} LN(P(g = i, X = j)), \tag{20}$$

where n_{ij} refers to the number of cases in each cell of cross-table of g and X and LN refers to natural logarithm taken during the process of each combination of i and j (see, Zaionts 2021).

One challenge in R_{PC} is that, using the established routines (e.g., Lancaster and Hamdan 1964; Martinson and Hamdan 1972; Olsson 1979), the estimates cannot reach the extreme values + 1 and –1, because the deterministic patterns lead to

computational problems. In the empirical section, R_{PC} is calculated manually using Zaiont's (2021) procedure with certain restrictions in the algorithm: first, a small positive number (10^{-7}) was added to each term that included logarithm as logarithm cannot be taken of a zero. Second, PMC embedded in the process cannot take the actual value 1.000, although a value close to 1, such as 0.9999999, can be allowed. Hence, technically, R_{PC} cannot reach the value 1 but it can be very close.

In traditional software packages such as IBM SPSS, the syntax for R_{PC} is not available, although some macros are (see Lorenzo-Seva and Ferrando 2015). In SAS, the command PROC CORR provides R_{PC}. Correspondingly, in RStudio, as an example, R_{PC} is calculated by *CorPolychor(x, y, ML = FALSE, control = list(), std.err = FALSE, maxcor = 0.9999)## S3 method for class 'CorPolychor' print(x, digits = max(3, getOption("digits")—3), …)* (see https://rdrr.io/cran/DescTools/man/CorPolychor.html).

2.2.3 r-bireg and r-polyreg correlation

In the early years of item analysis, the most used estimator of correlation between an item and score was biserial (R_{BS}) and polyserial (R_{PS}) correlation for estimating the inferred correlation between an observed variable (X) and an unobservable latent variable truncated to an ordinal form (g) (see Clemans 1958). However, even then, it was known that R_{BS} and R_{PS} tend to overestimate correlation in an obvious manner (R_{BS} and $R_{PS} > 1$) when ρ_{gX} is high to start with (see Footnote 1). Over the years, many solutions have been offered for this obvious challenge (see the history in Moses 2017). Maybe the best options, by far, is a coefficient called r-bireg and r-polyreg correlation (R_{REG}; see Livingston and Dorans 2004; Moses 2017).

Combining the notation by Livingston and Dorans (2004), Moses (2017), and the conceptualization above, the procedure assumes that the observed value in item g (x_i) is determined by an underlying latent continuous variable θ. The distribution of θ for test-takers with the observed value (y) in the score variable X reflecting θ is assumed to be normal with mean $= \beta y$ and variance $= 1$, where β is an item parameter estimated by the probit regression model $P(x_i \le 1|y) = \Phi(a_i - \beta_i y)$ where Φ is the standard normal cumulative distribution function and a_i and β_i are intercept and slope parameters. After the ML estimate of β is computed, R_{REG} is calculated as

$$\rho_{REG} = \frac{\beta \sigma_X}{\sqrt{\beta^2 \sigma_X^2 + 1}}, \tag{21}$$

where σ_X^2 is the population variance of the score variable X. The β-value can be calculated, for example, in IBM SPSS software using the syntax:

GENLIN g (ORDER = ASCENDING) WITH X/MODEL X
DISTRIBUTION = MULTINOMIAL
LINK = CUMPROBIT/CRITERIA METHOD = FISHER/PRINT SOLUTION.

2.2.4 G and G_2

Goodman–Kruskal G estimates the probability that observations in two variables are in the same order. G strictly reflects the (slightly modified) proportion of logically ordered test-takers by item after they are ordered by the score (Metsämuuronen 2021b). The computational form of G is usually expressed using the concepts of concordance and discordance between the observed values of pairs of test-takers in g (x_k, x_l) and X (y_k, y_l). If a pair of observations x_k and x_l and corresponding y_k and y_l have ranks in the same direction, the pair is concordant. Correspondingly, if the pair has the ranks in opposite order, the pair is discordant. Denoting the number of concordant pairs by P and the number of discordant pairs by Q, G proportions $P - Q$ to the number of pairs where the direction is known

$$G = \frac{P - Q}{P + Q}. \tag{22}$$

Notably, in this form, P and Q are twice of that of the simplified forms often seen in the textbooks (e.g., Metsämuuronen 2017; Siegel and Castellan 1988).

Traditionally, G has been taken as a symmetric measure, because it produces only one estimate the same manner as PMC (e.g., IBM 2017; Sheskin 2011; Sirkin 2006; Wholey et al. 2015). However, Metsämuuronen (2021b) showed that G has a hidden directional nature in the same manner and same direction as PMC and $D(g|X)$ have. When the scales of two variables are not identical, G is an unambiguously directional coefficient and the variable with a wider scale (X) explains the response pattern in the variable with a narrower scale (g), that is, "g given X" from the conditions viewpoint or "score dependent" as in generally known software packages.

Dimension-corrected G (G_2) is an estimator proposed by Metsämuuronen (2021a) seeing the deficiency in G to underestimate the correlation between item and score in an obvious manner when the number of categories in item exceeds four (see Metsämuuronen 2021a; see also later Fig. 7b). The computational form of G_2 is

$$G_2 = G \times (1 + (1 - abs(G)) \times A), \tag{23}$$

where G is the observed value of G and

$$A = \frac{df(g) - 1}{df(g)} \left(1 - \frac{1}{df(g)}\right)^2, \tag{24}$$

where $df(g)$ = (number of categories in the item – 1). Because of the cubic element A, G_2 has a "semi-trigonometric nature" (Metsämuuronen 2021b) in comparison with G, which has a strict linear nature (see later Fig. 6). When the scale of item has more than two categories, the magnitude of the estimates by G_2 tends to be higher in comparison with those by G except when the discrimination is deterministic ($G = G_2 = 1$). Notably, G_2 is not a general transformation but, instead, specific to measurement modelling settings where g and X are mechanically related (see discussion and warnings in Metsämuuronen 2021b).

In traditional software packages such as IBM SPSS, for instance, the syntax for G is CROSSTABS /TABLES = item BY Score /STATISTICS = GAMMA. In SAS, the command PROC FREQ provides G by specifying the TEST statement by GAMMA, SMDCR options. Correspondingly, in RStudio G is calculated by *GoodmanKruskalGamma*(*x, y = NULL, conf.level = NA, …*) (see https://rdrr.io/cran/DescTools/man/). For the empirical section, the estimates by G_2 are calculated manually based on the observed G and $df(g)$.

2.2.5 D and D_2

Somers' D is a close sibling of G; as G, D too estimates the probability that test-takers are in the same order in g and X, although the magnitude of the estimates by D are more conservative than those by G. This is caused by the fact that D proportions $P - Q$ with *all* possible pairs including the tied pairs. The computational form of D directed, so that X explains the response pattern in g, is

$$D(g|X) = D = \frac{P - Q}{D_g} = \frac{P - Q}{P + Q + 2T_g} \tag{25}$$

(IBM 2017; Metsämuuronen 2021b; see the rationale for the direction in Metsämuuronen 2020a, b), where $D_g = N^2 - \sum_{i=1}^{R} \left(n_{gi}^2 \right)$ refers to the number of all possible combinations of pairs and n_{gi} is the number of cases in the categories $g = i$, and $2T_g$ refers to the number of tied pairs related to g. By comparing forms (22) and (25), the reason for the conservative nature in D is obvious: while the estimates by G are not affected by the number of tied cases, the estimates by D are. Because of the connection to Jonckheere–Terpstra test statistics, D strictly reflects the proportion of logically ordered test-takers in g after they are ordered by X (Metsämuuronen 2021b).

Metsämuuronen (2020a) reminds us that D has a long history in the measurement modelling setting, although many has not recognized it (see also the history in Berry et al. 2018). Namely, Newson (2008) showed that Cureton's rank-biserial correlation (R_{RB}) is a special case of D. Therefore, Metsämuuronen (2021b) proposes that D could be called rank-*poly*serial correlation coefficient, because, while ρ_{RB} is restricted to binary items, D can be also used with polytomous items.

Dimension-corrected D (D_2) is an estimator proposed by Metsämuuronen (2020b) and corrected in Metsämuuronen (2021a) against the deficiency of D to underestimate the correlation between item and score in an obvious manner when the number of categories in item exceeds three (Metsämuuronen 2020a; see also Göktaş and İşçi 2011; see later Fig. 7b). The computational form of (corrected) D_2 is

$$D_2 = D \times (1 + (1 - abs(D)) \times A) \tag{26}$$

(Metsämuuronen 2021a), where D is the observed value of $D(g|X)$ and A is as Eq. (24). The magnitude of the estimates by D_2 tends to be higher in comparison with those by D except in two situations: when the discrimination is deterministic

($D = D_2 = 1$) and when the scale of g has just two categories causing $A = 0$. Like D and G, D_2 and G_2 also are close siblings. As with G_2, D_2 also has a semi-trigonometric nature and it is not a general transformation but specific to measurement modelling settings where g and X are mechanically related (see discussion and warnings in Metsämuuronen (2020b, 2021a).

In traditional software packages such as IBM SPSS, for instance, the syntax for D is CROSSTABS /TABLES = item BY Score /STATISTICS = D. In SAS, the command PROC FREQ provides D by specifying the TEST statement by D, SMDCR options. Correspondingly, in RStudio, D is calculated by *SomersDelta*(*x, y = NULL, direction = c("row", "column"), conf.level = NA, …)* (see https://rdrr.io/cran/DescTools/man/). For the empirical section, the estimates by D_2 are calculated manually based on the observed D and $df(g)$.

2.2.6 Attenuation-corrected PMC and eta

Metsämuuronen (2020c; see also 2022c) showed that PMC has a hidden directional nature, because its formula is shown to equal with a formula of a certain direction of the genuinely directional coefficient *eta* (Pearson 1903, 1905). Because PMC is known to be seriously attenuated when the scales of variables differ from each other as is usual in settings related to *eta* (in GLM settings) and item and score (in measurement modelling settings), Metsämuuronen (2022c) suggests that both *Rit* and *eta* should be attenuation-corrected.

Attenuation correction in PMC has been studied from Pearson (1903) and Spearman (1904) onwards. The traditional corrections (see the mechanics in, e.g., Sackett and Yang 2000; Schmidt et al. 2008) are based on correcting restriction in range when restriction of range has occurred, typically, in the score variable. In measurement modelling settings and in settings related to *eta*, this approach does not seem to be the best option, because the attenuation happens *between* an item and the score and not in the score alone. Notably, the traditional procedures of calculating *eta* give us only the positive values of the coefficient (see Metsämuuronen 2022c). Hence, before the attenuation correction for *eta*, the correct value of *eta*—including the negative values also—is preferable to be used. Correction has, however, no effect on eta squared which is usually used in the settings familiar from general linear modelling. It may have, though, a notable effect on the estimates by eta itself and, consequently on the interpretation of the estimates. For this, a corrected form of eta (with binary items, $\eta(g|X) = \left(\overline{X}_{X1} - \overline{X}_{X0}\right) \times \frac{\sigma_g}{\sigma_X}$ where $\overline{X}_{X1}$ and $\overline{X}_{X0}$ are the scores in the subpopulations $g = 0$ and $g = 1$, σ_g refers to the standard deviation of g, and σ_X is the standard deviation of X) or a simple transformation $sign(Rit) \times eta$ (with polytomous items) could be used (see the derivations and rationales in Metsämuuronen 2022c).

Metsämuuronen (2022b, c) suggests a simple attenuation correction to *Rit* (ρ_{AC}, R_{AC}) as the proportion of the observed item–score correlation (ρ_{gX}^{Obs}) of maximal correlation (ρ_{gX}^{Max}) possible to obtain with the observed g and X

$$\rho_{AC} = \frac{\rho_{gX}^{Obs}}{\rho_{gX}^{Max}} = \frac{\rho_{gX}}{\rho_{gX}^{Max}}. \tag{27}$$

Similarly, the attenuation-corrected η (η_{AC}, E_{AC}) is the proportion of observed *eta* ($\eta_{g|X}^{Obs}$; see the discussion of correct direction in Metsämuuronen 2020a, 2022c) and the maximal *eta* ($\eta_{g|X}^{Max}$) possible to obtain given the variance of the score

$$\eta_{AC} = \frac{\eta_{g|X}^{Obs}}{\eta_{g|X}^{Max}} = \frac{\eta(g|X)}{\eta_{g|X}^{Max}}. \tag{28}$$

The maximum values of both *Rit* and *eta* in the given dataset are obtained when the correlation is calculated between the *independently* ordered variables g and X; this maximizes the item–score covariance (see Eq. 19) needed in maximizing *Rit* (see Metsämuuronen 2022c) and minimizes the element $SS_{Error} = \sum \left(y_{ij} - \overline{X}_{Xg} \right)^2$ needed to maximize *eta* (see Metsämuuronen 2022c). Except in the special case where the variables are in the same order and, hence, have reached the maximal possible value leading to $\rho_{gX}^{Obs} = \rho_{gX}^{Max}$ and $\eta_{g|X}^{Obs} = \eta_{g|X}^{Max}$, $R_{AC} > Rit$ and $E_{AC} > eta$. Otherwise, the characteristics of R_{AC} and E_{AC} are largely unstudied. The latent linear or trigonometric nature of R_{AC} is ambiguous: if we interpret Eq. (27) to be a linear transformation of *Rit*, the trigonometric nature of *Rit* is inherited to R_{AC}, while, if we interpret R_{AC} as a proportion of the maximal value, the outcome may come close the linear nature embedded in G, D, and *tau-b*.

In the traditional software packages such as IBM SPSS, for instance, the syntax for *eta* is CROSSTABS /TABLES = item BY Score /STATISTICS = ETA. In SAS, the positive values of *eta* can be found by taking square of eta squared after PROC GLM with option EFFECTSIZE. Correspondingly, in RStudio, eta is calculated by *eta(x, y, breaks = NULL, na.rm = FALSE)* (see https://rdrr.io/cran/ryouready/man/eta.html). For the maximal *Rit* and *eta*, in R, the variables (vectors) can be sorted by a command *sort (x) #*. For the empirical section, R_{AC} and E_{AC} were calculated manually. For correct values of *eta* also including the negative values, a simple transformation $sign(Rit) \times eta$ suggested by Metsämuuronen (2022c) is used.

2.2.7 Kendall *tau-b*

Kendall *tau-b* (Kendall 1948) belongs to the same family as G and D: with continuous variables, G, D, and *tau-b* equal *tau-a* (Kendall 1938). As G and D, *tau-b* estimates the probability that test-takers are in the same order in g and X. In comparison with G and D, *tau-b* is a truly symmetric measure. Using the same notation as with G and D

$$tau\text{–}b = \frac{P - Q}{\sqrt{D_g \times D_X}} = \frac{P - Q}{\sqrt{(P+Q)^2 + 2(P+Q)\left(T_g + T_X\right) + 4\left(T_g \times T_X\right)}}, \tag{29}$$

(e.g., IBM 2017; Metsämuuronen 2021b) where $D_X = N^2 - \sum_{j=1}^{C} \left(n_{Xj}^2 \right)$, n_{Xj} is the number of cases in the categories $X=j$, and T_X refers to the tied pairs related to X. By comparing the forms of D, G, and *tau-b*, the lowest magnitude of the estimates is expected from *tau-b* due to extensive use of tied pairs.

In traditional software packages in calculation such as IBM SPSS, for instance, the syntax for *tau-b* is CROSSTABS/TABLES = item BY Score /STATISTICS = TAU. In SAS, the command PROC CORR produces tau-b. Correspondingly, in RStudio, *tau-b* is estimated by either `cor(..., method="kendall")` or *KendallTau-b*(*x, y = NULL, conf.level = NA, …)* (see https://rdrr.io/cran/DescTools/man/).

2.3 Criteria and thresholds for the evaluation

In what follows, in all three sub-studies, a rough, simple, and partly subjective mechanism of scoring is used to evaluate the magnitudes of artificial mechanical error in the estimates by different estimators of correlation. The rough method makes sense, because the multiple criteria differ notably from each other, the factual magnitude of the mechanical error in estimation depends on several factors with varying magnitude, and because the more nuanced, standardized, methods were not available for this unifying treatment. The logic in scoring is condensed in Table 1 and discussed below. Obviously, because of based on rough and partly subjective boundaries, another evaluator could set standards to different levels and rank the estimators partly differently.

In Sub-study 1 in Sect. 3, seven sources of error are studied to evaluate the magnitude of artificial mechanical error in selected estimators of correlation. A five-point ordinal scale is in use. If the estimator shows *no effect* by a specific source of MEC, + 2 is given. For example, with R_{PC} and G with source (2) of MEC referring to item difficulty, the estimates by R_{PC} and G systematically detect the perfect latent correlation regardless of the item difficulty and, hence, + 2. In the other extreme, if the estimator is "remarkably" affected by the source of MEC lowering the estimate, − 2 is given. For example, with *tau-a* and *Rit* with source (2), the estimate may vary from 0 to 0.87 depending on item difficulty regardless of the perfect latent correlation. This is taken as a "remarkable" error in estimation. Notably, we could have found estimators that would underestimate the latent correlation even more drastically. Those would be given the value − 2 also. Less extreme scores − 1 and + 1 are given if *some* error, although not a particularly notable one, is detected (− 1), or if the estimates are very close to the best options although still having some error (+ 1). For example, with the source (2), D and D_2 show slight underestimation depending on item difficulty and, hence, + 1. With an unknown effect, 0 is given.

In Sub-study 2 in Sect. 4, the scale for the scoring scheme is − 1 to + 1. It is known that the trigonometric and directional nature of the coefficient leads to higher approximations of correlation and this is taken as a positive matter and, hence, + 1 in the scoring systemic. In contrast, linear and symmetric nature of the coefficient

Table 1 Scoring scheme for the magnitude of mechanical underestimation of coefficients of correlation

Source of MEC	*–2* *"remarkably affected"*	*–1* *"notably affected"*	*0* *"unknown effect"*	*+1* *"slightly affected"*	*+2* *"not at all affected"*
	Effect in units of correlation				
Sub-study1					
(1) Discrepancy of scales	> 0.2	0.1–0.2	Unknown	< 0.1	0
(2) Item difficulty and variance	> 0.6	0.2–0.6	Unknown	< 0.2	0
(3) Distribution of the latent variable	> 0.2	0.1–0.2	Unknown	< 0.1	0
(4) Number of categories in the item	> 0.2	0.1–0.2	Unknown	< 0.1	0
(5) Number of categories in the score	> 0.1	0.05–0.1	Unknown	< 0.05	0
(6) Number of items forming the score	> 0.1	0.05–0.1	Unknown	< 0.05	0
(7) Number of tied cases in the score	> 0.1	0.05–0.1	Unknown	< 0.05	0
Sub-study 2					
(8) Linear or trigonometric nature	–	Linear	Unknown	Trigonometric	–
(9) The directional or symmetric nature	–	Symmetric	Unknown	Directional	–
Sub-study 3					
(10) Instability in estimates in reflecting population parameter: maximal difference between the estimates related to varying *df*(*g*)	–	Notable general instability (> 0.016)	Instable only with very small samples (0.002–0.016)	Not notable instability (< 0.002)	–
(11) possible overestimation	–	Systematic overestimation > 0.01	Very small, unsystematic overestimation < 0.01	Under-estimation (no tendency for overestimation)	–

leads to lower approximations of correlation and this is taken as a negative matter and, hence, − 1 in the scoring system. Again, with an unknown effect, 0 is given.

In Sub-study 3 in Sect. 5, the scale is, again, − 1 to + 1 although subjectivity is somewhat higher than in Sub-studies 1 and 2. Sub-study 3 studies the instability and overestimation in relation to a real-world dataset. Evaluating the magnitude of "instability" is subjective. However, some rough boundaries are used in evaluation: if the estimators show notable instability between the population value and the estimates − 1 is given. If the average of the estimates for items with binary scale on one hand and items with wide scale (11–16 categories) on the other differ more than 0.016 units of correlation this is considered as "notable" differences (− 1). Similarly, a difference of a round 0.002 units of correlation is taken as a small effect (+ 1). When it comes to overestimation, a systematic overestimation of size of 0.01 units of correlation was taken "notable" (− 1). Because of the basic principle of being merely too tight in statistical inferences, the condition of no tendency for overestimation means, in practice, that a condition of *underestimation* of the population correlation is taken as a positive matter (+ 1).

3 Study 1. General characteristics of estimators of correlation to reach the perfect population correlation

3.1 Research question in Study 1

Study 1 examines the extent to which different estimators of correlation reflect the true correlation between two variables under the condition specific to measurement modelling settings that a common latent variable θ drives both item and score causing the true correlation between the item and the score to be perfect. What of interest is, specifically, in seven first sources of MEC: sensitivity to (1) discrepancy of scales, (2) item difficulty and variance, (3) distribution of the latent variable, (4) number of categories in the item, (5) number of categories in the score, (6) number of items forming the score, and (7) number of tied cases. The behavior of *Rit* is compared with various alternative estimators by varying the latent variables (normal, skewed normal, and uniform), the degrees of freedom of $df(X) = C - 1$, $df(g) = R - 1$, and the difficulty level of g (p).[2]

3.2 Datasets used in Study 1

For Study 1, three vectors with $N = 1000$ cases were formed: a standardized normal vector with $N(0,1)$, a skewed-normal vector with $\Gamma(2,1)$, and a uniform vector without tied cases. The last was simply a variable with values 1–1000 in a consecutive order. Each vector was duplicated to form three pairs of (perfectly correlated) variables. Each pair of vectors was manipulated, so that one of the identical vectors

[2] The degrees of freedom, like what is used with chi squared statistic, is a relevant statistic here, because the analysis is done using two-way contingency tables of the variables.

became a variable with a narrower scale (item g) and the other with a wider scale (score X). By changing the cut-off of the original vector, the scale of X related to the normal and gamma distributions was set to vary with $df(X) = 4$, 6, 12, 20, 25, 30, 40, and 60 and the uniform distribution with $df(X) = 4$, 9, 19, 24, 39, 49, and 99.The dfs in the set of uniform distribution were selected, so that all the categories would have equal number of cases, that is, for example, when $df(X) = 4$, there are five categories (0–4 or 1–5), and 1000/5 leads us to 250 cases in each consecutive category in X. Similarly, $df(X) = 99$ leads to 100 categories with 1000/100 = 10 cases in each category. The scale of g was set to vary with fixed values $df(g) = 1$, 2, 3, and 4, that is, the most commonly used scales from a binary to a 5-point Likert type of scales were covered. Item difficulty was varied by changing systematically the cut-off for the bins.

The dataset comprising 22,824 estimates by each estimator of interest was formed by the following steps:

(1) A standardized normal vector with $N(0,1)$, a skewed-normal vector with $\Gamma(2,1)$, and a uniform vector without tied cases were formed and duplicated.
(2) Of the normal- and gamma-distributed original vectors, eight score variables ($df(X) = 4$, 6, 12, 20, 25, 30, 40, and 60) were formed by multiplying the original vector and cutting systematically the original vector of 1000 cases into 5, 7, 13, 21, 26, 31, 41, and 61 categories, so that always the form is either normal or gamma-distributed. The uniform vector was multiplied and cut into seven score variables ($df(X) = 4$, 9, 19, 24, 39, 49, and 99), that is, the original 1000 cases were cut into 5, 10, 20, 25, 40, 50, and 100 categories.
(3) The other version of identical vectors of normal-, gamma- and uniform-distributed variables formed the items with fewer categories than in the score version of the vector. The binary variables were formed first. 1000 cases can be cut into 250–1 categories by systematically increasing the cut-off by four cases starting from the highest scoring cases so that 4 highest-ranked cases out of 1000 were given 1 s and the rest 996 cases were given 0 s, i.e., $p = 4/1000 = 0.004$, $8/1000 = 0.008$, $12/1000 = 0.012$, and so on up to $p = 0.996$. This logic was used for the gamma and uniform distributions: 249 items were formed with increasing difficulty levels. For the normal distribution, the logic was different: binning was based on how many cases would be selected in each bin if a truly normal provision would be used. The most extreme item was the one with $p = 0.002$. From this on, the items were formed by an increment of one case, that is, the item difficulty was $p = 0.003$, 0,004, … up to $p = 0.030$. After this, the cases were selected by the increment of an uneven number of cases leading to $p = 0.030$, 0.032, 0.034 … up to $p = 0.986$, 0.997, and 0.998. From the normal-distributed vector, 243 items were formed. Altogether, 249 (items) × (8 + 7) (scores) + 243 (items) × 8 (scores) = 5,679 estimates of binary items vs. score variables with varying number of categories were formed for each coefficient of correlation of interest.
(4) The items with three, four, and five categories were formed using the same logic as with binary items: the cut-offs for binning the gamma and uniformly distributed vectors were done systematically, so that 249 variables were formed

Table 2 Effects of seven sources of MEC to the estimates

Source of MEC:	*Rit*	R_{PC}	R_{REG}	G	D	G_2	D_2	*tau-b*	R_{AC}	E_{AC}
(1) Discrepancy of scales	− 2	+2	+2	+2	+1	+2	+1	− 2	+2	+2
(2) Item difficulty and variance	− 2	+2	+2	+2	+1	+2	+1	− 2	+2	+2
(3) Distribution of the latent variable	− 2	+2	+2	+2	− 2	+2	− 2	+1	+2	+2
(4) Number of categories in the item	− 2	+2	+2	+2	+1	+2	+1	− 2	+2	+2
(5) Number of categories in the score	− 1	+2	+1	+2	− 2	+2	− 2	− 1	+2	+2
(6) Number of items forming the score	− 1	+2	+1	+2	− 2	+2	− 2	− 1	+2	+2
(7) Number of tied cases in the score	− 1	+2	+1	+2	− 1	+2	− 1	− 2	+2	+2
SUM	− 11	+14	+11	+14	− 4	+14	− 4	− 9	+14	+14

scale: +2 = *no effect* = *MEC-free*, +1 = *insignificant effect*, 0 = *unknown effect*, − 1 = *notable effect*, − 2 = *remarkable effect lowering the estimate*

with systematically varying difficulty levels. Notably, then, the difficulty levels (p values) are not as systematically increasing as in the binary case. For the normal vector with three categories, 265 items were formed and for three and four categories 344 items were formed. In combining items and scores, notably, not all combinations were used. For example, with 5 categories in an item, the shortest relevant score would be of 10 categories (formed of two items) instead of 6 or 8 categories which were, however, relevant for the binary case.

Finally, the dataset consists of 5679 estimates of binary items vs. score variables, 5855 estimates of three-category items vs. score variables, 5645 estimates of four-category items vs. score variables, and 5645 estimates of five-category items vs. score variables, that is, altogether 22,824 estimates for all estimators of correlation of interest in Study 1. This dataset is available in SPSS format at http://dx.doi.org/10.13140/RG.2.2.20111.30882 and in CSV format at http://dx.doi.org/10.13140/RG.2.2.17241.65127.

3.3 Main results from Study 1

Table 2 summarizes the results of seven first sources of MEC.

The main result of the simulation is that, of the estimators in the comparison, R_{PC}, G, G_2, R_{AC}, and E_{AC} are not affected by *any* of the seven sources of MEC *at all*; in all seven conditions, they correctly produce the ultimate value $G = G_2 = R_{AC} = E_{AC} = 1 \approx R_{PC}$ indicating that the error related to MEC is zero ($e_{wi\theta\_\mathrm{MEC}} = 0$; see Sect. 1.3). Of the coefficients in comparison, *tau-b* and *Rit* suffer the most of the seven sources. The impact and mechanism of the sources are discussed below.

3.4 Effects of the discrepancy between the scales, the difficulty level of the item, the number of categories in the item, and latent variable

Of the estimators in comparison, *Rit* and, specifically, *tau-b* are remarkably affected by the discrepancy between the scales, the difficulty level of the item,

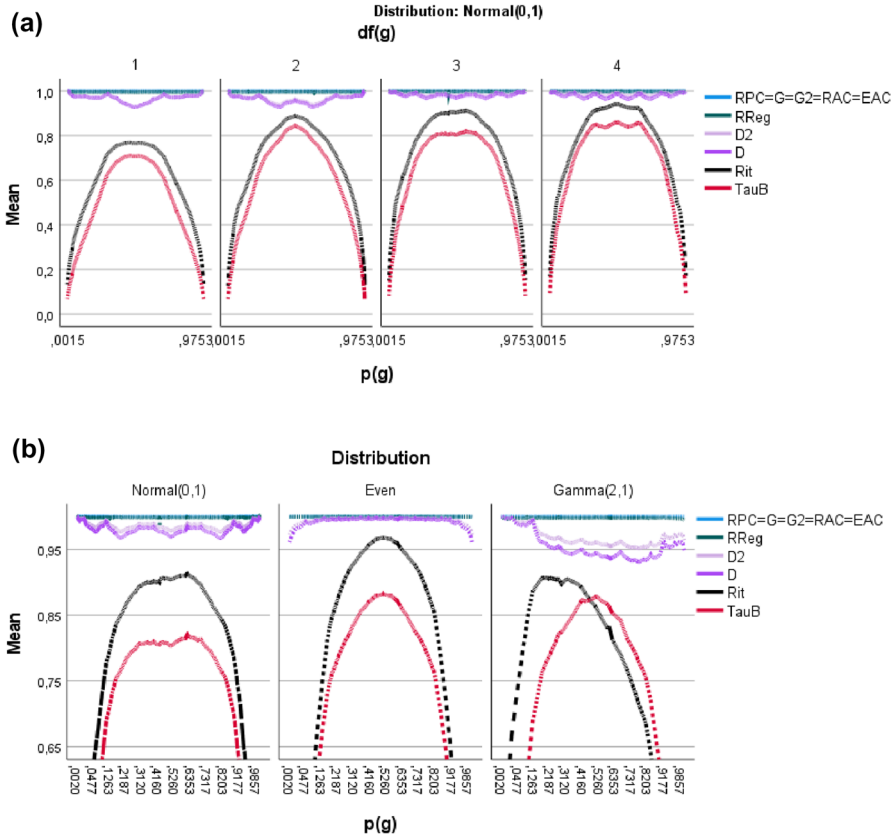

Fig. 3 **a** Effects of scales, item difficulty, and *df*(*g*) in the item on the estimates. Note: *p*(*g*) = item difficulty, Mean = estimate of correlation. **b** Effects of scales, item difficulty, and latent distribution on the estimates with *df*(*g*) = 2. Note the cut scale in *Y* axis. Note2: *p*(*g*) = item difficulty. Mean = estimate of correlation

and the number of categories in the item (see Fig. 3a and b more details in Appendix 1) and, hence, − 2 in Table 1. While *Rit* is remarkably affected also by the latent distribution, *tau-b* seems to be less affected by this source; in the latter, the maximum estimate does not depend on the latent distribution, although the widths of the curves differ to some extent. Hence, + 1 for *tau-b* in Table 1. In comparison with *Rit*, the estimates by *D* and R_{REG} are less affected by MEC although, of the two, *D* is mildly more affected by item difficulty (+ 1) and notably more by the latent distribution (− 2), while in R_{REG}, the loss of information is nominal in this regard (+ 2). Also, *D* and D_2 are affected by the number of categories (− 2), while its effect in R_{REG} is more nominal although real (+ 1). R_{PC}, *G*, G_2, R_{AC}, and E_{AC} are not affected at all by this (or these) specific source of MEC (and, hence, + 2 in Table 1).

Table 3 Variables with perfect latent correlation forming 3 × 13 cross-table

		X													Total
		0	1	2	3	4	5	6	7	8	9	10	11	12	
g	0	3	9	26	65	121	84	0	0	0	0	0	0	0	308
	1	0	0	0	0	0	92	200	92	0	0	0	0	0	384
	2	0	0	0	0	0	0	0	84	121	65	26	9	3	308
Total		3	9	26	65	121	176	200	176	121	65	26	9	3	1000

3.5 Effect of the number of tied cases, categories in X, and items in the test

To illustrate the *effect of the number of tied cases* in the estimators, let us consider a pair of variables with $df(g) = 2$ and $df(X) = 12$ with latent normality (Table 3); this illustrates the calculation of the estimates also. This effect is seen, specifically, with short tests causing more tied values in the score (see Fig. 4).

Given Table 3, $Rit = 0.878$, and $R_{PC} \approx 1.000$; the low value in the former is partly caused by the tied cases, and the perfect value in the latter is caused by the fact that the variables are in the same order. For R_{REG}, the magnitude of β is $\hat{\beta} = 6.762$ and $\hat{\sigma}_X^2 = 3.972$. Hence, $\hat{\rho}_{REG} = \frac{6.762 \times 1.993}{\sqrt{6.762^2 \times 3.972 + 1}} = 0.997$, that is, R_{REG} loses some information, but the loss is nominal in magnitude. For *G, D, G*$_2$*, D*$_2$*,* and *tau-b,* $P = 2 \times [(3 + 9 + \ldots + 121) \times (384 + 308) + \ldots + 92 \times (121 + \ldots + 3)] = 631{,}904$, $Q = 0$, $D_g = 1000^2 - (2 \times 308^2 + 384^2) = 662{,}816$, $D_X = 1000^2 - (3^2 + \ldots + 3^2) = 858{,}784$, and, because $df(g) = 2$, $A = 0.5 \times 0.25 = 0.125$. These lead to $G = G_2 = (P - Q)/(P + Q) = P/P = 1.000$, $D = (P - Q)/Dg = 631{,}904/663{,}816 = 0.953$, $D_2 = 1 - (D - 1) \times (A - 1) = 1 - (0.953 - 1) \times (0.125 - 1) = 0.959$, and $tau\text{-}b = (P - Q)/(\sqrt{D_g \times D_X}) = 631{,}904/\sqrt{663{,}816 \times 858{,}784} = 0.838$. For R_{AC} and E_{AC}, because the item and score are in the same order, *Rit* and *eta* have reached the maximal possible value leading to $\rho_{gX}^{Obs} = \rho_{gX}^{Max}$ and $\eta_{g|X}^{Obs} = \eta_{g|X}^{Max}$ and, consequently, $RAC = \rho_{gX}^{Obs} \Big/ \rho_{gX}^{Max} = 1.000$ and $EAC = \eta_{gX}^{Obs} \Big/ \eta_{gX}^{Max} = 1.000$.

To outline the effect of the tied cases, the effect is obvious when we compare the forms of *D*, *G*, and *tau-b* (see also later Fig. 7b): because *tau-b* extensively uses tied pairs in calculation, it is remarkably affected by tied pairs (–2), *D* and *D*$_2$ are less affected (–1), and because *G* and *G*$_2$ omit the tied pairs, they are not affected by the number of tied pairs at all (+2). From the viewpoint of magnitude of the estimates, we may infer that R_{PC} is not affected of the tied cases at all (+2), while *Rit* seems to be affected in some extent (–1), and, although we do not know the exact effect, the effect seems to be only nominal in R_{REG} (+1).

The effect of the *number of categories in X* is illustrated in set of graphs in Fig. 5 (see more details in Appendix 2). In real-life test settings, the number of categories in the score is connected to *number of tied cases* and *number of items in the test* as well: the small number of categories indicates that the number of items is small, and the less categories in the score and the more test-takers, the more tied cases. Hence, the effect of the number of categories in *X* (source 5 in Table 2) is not independent of source 6 and, therefore, their effect is evaluated

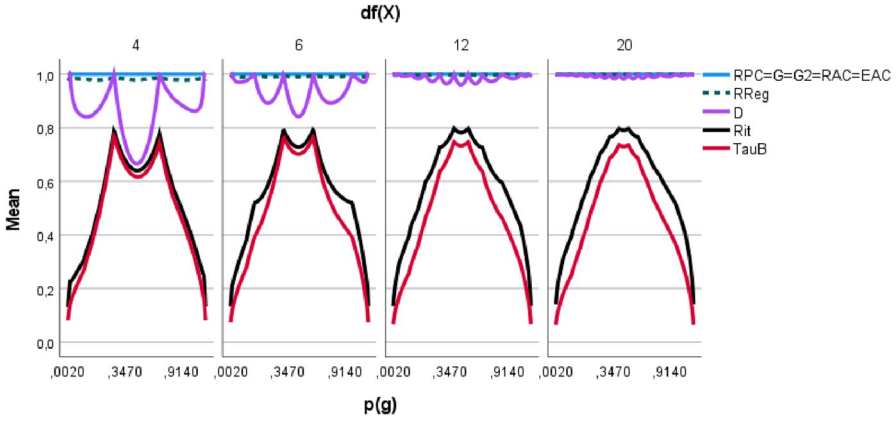

Fig. 4 Effect of tied cases with short tests ($df(X) \leq 20$) and binary items ($df(g) = 1$)

as identical in Table 2. Again, R_{PC}, G, G_2, R_{AC}, and E_{AC} are not affected at all by these specific sources of MEC (and, hence, +2 in Table 2). *Rit*, D, D_2, and R_{REG} are affected by the number of categories in the score in some extent, and the patterns are mainly similar: if there are 12 categories in the score or less, the loss of information is notable, while, if there are 20 or 25 categories or more, the loss of information does not increase (see also Fig. 3 above). The effect is relatively small in R_{REG} (+1), although the pattern follows the same as seen with D and D_2. Figure 5 illustrates how D and D_2 (−2) are more affected by $df(X)$ than *Rit* (−1). *Tau-b* (−1) is also affected by $df(X)$ although by a different pattern: the more categories in X the smaller the magnitude of the estimates (see Appendix 2).

4 Study 2: underestimation of correlation in the real-world dataset

4.1 Research question in Study 2

Study 2 examines the underestimation of correlation caused by the linear or trigonometric nature of the estimator (source 8) and the directional or symmetric nature of the estimator (source 9). Source 8 is first considered from the theoretical viewpoint by connecting the estimators with Greiner's relation after which an empirical dataset is used to study the phenomenon with real-world items. Source 9 is studied for those estimators whose directionality is not known.

4.2 Datasets used in the Study 2

For Study 2, a real-world dataset is used to study the underestimation of the correlation under the condition that there are measurement errors in the variables. This

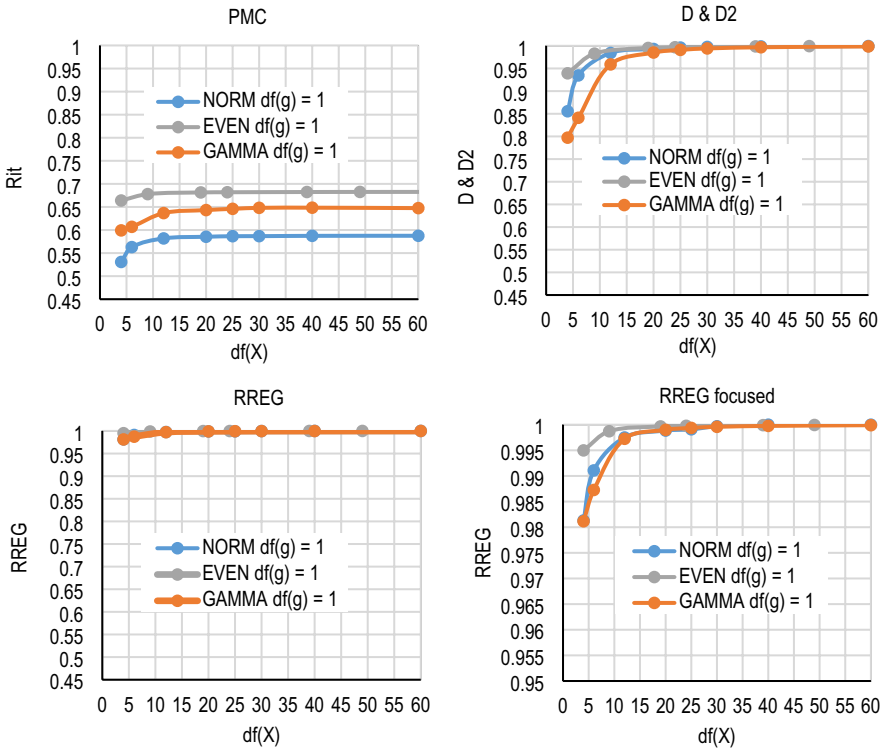

Fig. 5 Effect of the number of categories in the score and the latent distribution, mean of the estimates [$df(g) = 1$, $k = 22{,}824$ items]

dataset is based on a national-level dataset of 4,022 test-takers of a mathematics test with 30 binary items (FINEEC 2018). In the original dataset, the lower bound of the reliability was $\alpha = 0.885$, item discrimination ranged $0.332 < Rit < 0.627$ with the average $\overline{Rit} = 0.481$, and the difficulty levels of the items ranged $0.24 < p < 0.95$ with the average $\overline{p} = 0.63$. Ten random samples of $n = 25$, 50, 100, and 200 test-takers were drawn from the original dataset imitating different sizes of finite sample sizes typical to real-life testing settings: $n = 25$ may be a typical sample size in the classroom testing and $n = 200$ may be the sample size for a lecture for large student group or a common test in a school for all students of the same age. In each of the 40 datasets, 36 shorter tests were produced by varying the number of items, difficulty levels of the items, $df(g)$, and $df(X)$. The polytomous items were constructed as sums of the original binary items. As a result, the dataset consisted of 14,880 partly related test items from 1440 tests with a varying number of test-takers ($n = 25$, 50, 100, and 200), items ($k = 2$–30, $\overline{k} = 10.33$, std. dev. 8.621), lower bound of reliabilities ($\rho_\alpha = 0.55$–0.93, $\overline{\rho}_\alpha = 0.850$, std. dev. 0.049, the average difficulty levels ($\overline{p} = 0.50$–0.76, $\overline{p} = 0.66$. std. dev. 0.052), and ($df(g) = 1$–14, $\overline{df(g)} = 4.57$, std. dev. 3.480), and ($df(X) = 10$–27, $\overline{df(X)} = 18.06$, std. dev. 3.908). Average estimates are presented in Appendix 3. This dataset is

available in SPSS format at http://dx.doi.org/10.13140/RG.2.2.17594.72641 and in CSV format at http://dx.doi.org/10.13140/RG.2.2.10530.76482.

4.3 Main results from Study 2

To outline the analysis of Study 2, Table 4 summarizes the results concerning the sources 8 and 9 of MEC. Like in Study 1, a simple mechanism of ranking the estimators was used, however, with a reduced scale. If the estimator showed trigonometric or directional nature referring to the lower quantity of MEC, + 1 was given, if the latent nature was unknown, 0 was given, and, if the estimator showed linear or symmetric nature referring to higher quantity of MEC, − 1 was given.

Of the estimators in comparison, *G*, *D*, and *tau-b* underestimate the correlation between item and score in an obvious manner, caused by their linear nature, when the number of categories gets higher than 3 or 4—this is already known from the previous simulations (e.g., Metsämuuronen 2020b, 2021a). Without any real-world dataset, it is known that, except *tau-b*, all estimators in the comparison are either (positively) directional in their nature and, hence, they are logical from the testing theory viewpoint, or their directional nature is not known (R_{PC}, R_{REG}). The latter are studied initially using the real-world dataset. The impact and mechanism of the sources are discussed below.

4.4 Effect of the linear nature to the estimator

G is an interesting benchmarking estimator for the effect of linearity in the estimates. Although *G* accurately reflects the perfect latent correlation under all previous conditions related to the sources of MEC (see Study 1), it tends to underestimate the correlation in the real-life datasets in an obvious manner when the number of categories exceeds 4 (see Metsämuuronen 2021a), and hence, (− 1) in Table 4. This underestimation can be explained by Greiner's relation (Greiner 1909) discussed by Kendall (1949), Newson (2002), and Metsämuuronen (2020b, 2021a). Namely, with continuous variables *X* and *Y* (implying no tied pairs), $G = D = tau\text{-}b = tau\text{-}a$. Then, Greiner's relation states that

$$\rho_{XY} = \sin\left(\frac{\pi}{2} \times tau\text{-}a_{XY}\right) = \sin\left(\frac{\pi}{2} \times D_{XY}\right) = \sin\left(\frac{\pi}{2} \times G_{XY}\right) = \sin\left(\frac{\pi}{2} \times tau\text{-}b_{XY}\right). \tag{30}$$

From Eq. (30), it is known that, in the case of two continuous variables, except for the extreme values ± 1 and 0, the magnitude of the estimates by PMC tends to be greater than that by *G*, *D,* and *tau-b:* for $G = D = tau\text{-}b = 0.5$ we would expect to see $\text{PMC} = 1 \Big/ \sqrt{2} = 0.7071$. The trigonometric vs. linear nature of these coefficients is obvious when we plot the estimates in the same graph (Fig. 6).

Graphs in Fig. 7 illustrate the differences between the estimators regarding their linear or trigonometric nature. All in all, the traditional estimators *D*, *G*, and *tau-b* based on probability are prone to underestimate the correlation of a score and an item with a wide scale (and, hence, − 1 in Table 4). However, in binary settings, the score would

Table 4 Effects of sources 8 and 9 of MEC to the estimates

Source of MEC:	PMC	R_{PC}	R_{REG}	G	D	G_2	D_2	*tau-b*	R_{AC}	E_{AC}
(8) Linear or trigonometric nature[1]	+1	+1	+1	−1	−1	+1	+1	−1	+1	+1
(9) The directional or symmetric nature[1]	+1	±0	±0	+1	+1	+1	+1	−1	+1	+1
SUM	+2	+1	+1	±0	±0	+2	+2	−2	+2	+2

(1) scale: +1 = trigonometric/directional nature referring to lower magnitude of MEC, ±*0* = unknown nature, and − 1 = linear/symmetric nature referring to higher magnitude of MEC

be + 1. D_2 and G_2 (+1) with their semi-trigonometric nature are developed to overcome this obvious deficiency in D and G, and PMC and R_{PC} are clearly trigonometric in their nature (+1). Also, R_{AC} and E_{AC} seem to have inherited the trigonometric nature from PMC (+1); after all, the forms of *eta* and *Rit* are closely related (see Metsämuuronen 2020c, 2022c). The nature of R_{REG} is unknown in this regard in Table 4. However, the average estimates by R_{PC}, R_{REG}, and D_2 are almost identical which may indicate that also R_{REG} has, factually, latent trigonometric nature and hence (+1 in Table 4). The similarity in tendencies of R_{PC} and G_2 (with $df(g) < 4$) and R_{PC} and D_2 (with $df(g) > 3$) is an interesting phenomenon considering that R_{PC} reflects unreachable, theoretical constructs while G_2 and D_2 refer to observed variables.

4.5 Effect of the directional nature to the estimator

When it comes to the directionality of the estimators, the testing theory postulates that θ manifested as (the weighted or unweighted) X explains the behavior in the test item and not the other way (e.g., Byrne 2016; Metsämuuronen 2017). Hence, this direction in correlation makes sense in the measurement modelling settings. From the magnitude of the estimates by *tau-b* in comparison with D and G, we may infer that when the estimator reflects a genuine symmetric correlation, it tends to *under*estimate the actual correlation between g and X, while the directional estimators seem to *not over*estimate the correlation. Hence, the directional nature of the estimators can be taken as positive (+1) and the symmetric nature as negative (− 1) regarding MEC.

Of the estimators in comparison, *tau-b* is obviously a symmetric measure (− 1) and D is obviously a directional measure (+1); here, the direction selected for D was relevant from the item analysis settings viewpoint (see Metsämuuronen 2020a, 2021a). Selecting this direction also for *eta* (see Metsämuuronen 2020a, 2022c) leads to conclude that the direction selected for E_{AC} is also relevant from the item analysis settings viewpoint, hence, (+1). Also, PMC (+1) as well as G (+1) are found to have a hidden directional nature (see Metsämuuronen 2020c, 2021b). Consequently, D_2, G_2, and R_{AC} are directional measures (+1). The symmetric or directional nature of R_{PC} and R_{REG} is unknown and, hence, they are given 0 in Table 4. However, their behavior in the real-world dataset and, specifically, in relation to the different directions of coefficient *eta* hints that they also may have a hidden directional nature when the number of categories is high; if the item has five categories or more ($df(g) > 3$), the magnitude of the estimates by R_{PC} and R_{REG} tends to be closer to the estimates by coefficient *eta* directed to the same direction as *Rit* (see Fig. 8). Algebraic reasons for the close connection of

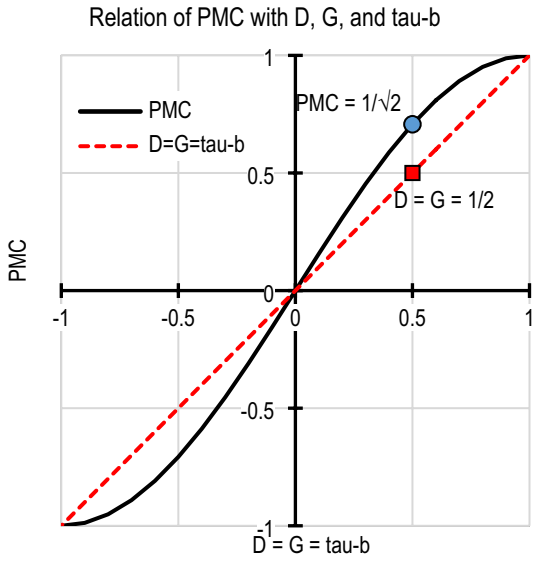

Fig. 6 Greiner's relation of PMC with *D*, *G*, and *tau-b* with continuous variables

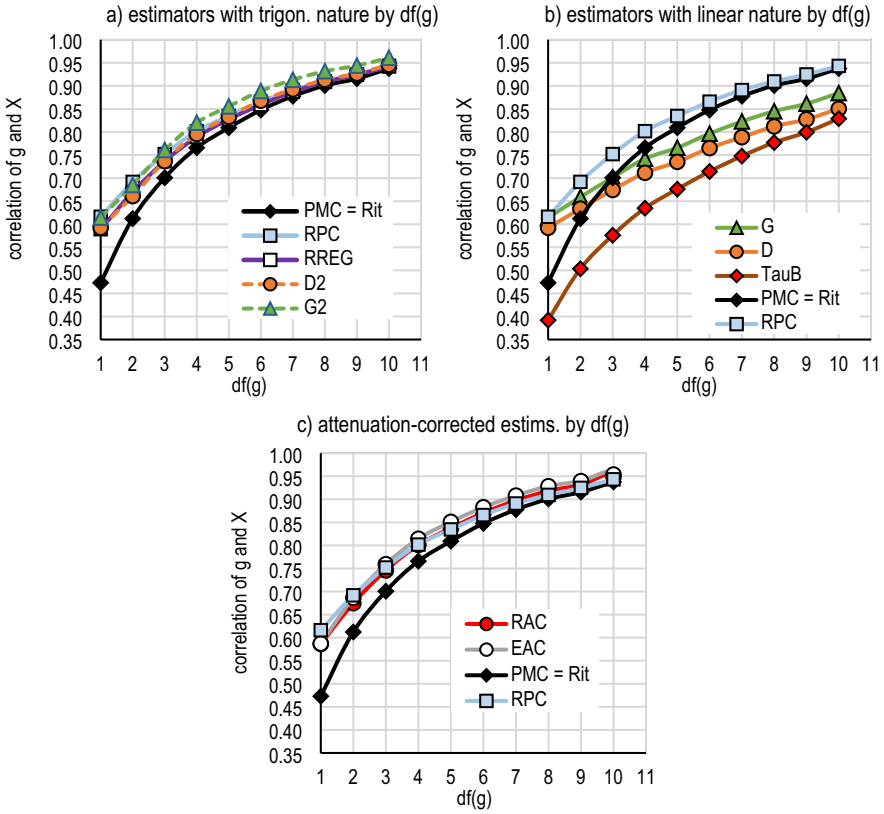

Fig. 7 Estimators with linear, trigonometric, and semi-trigonometric nature

Rit and *eta* directed, so that "*g* given *X*" or "*X* dependent" are discussed by Metsämuuronen (2022c).

5 Study 3: possible overestimation in the estimators of correlation

5.1 Research question in Study 3

Study 3 examines the instability of the estimators in reflecting the population correlation (Source 10) and possible overestimation of population correlation of the selected estimators (Source 11). Obviously, if the estimates are instable, we cannot trust the estimate. Also, both under- and overestimation are not optimal conditions. However, usually, the overestimation is a more negative alternative from the viewpoint of conservativeness in statistical inference. The question is: To what extent the sample estimates correspond with the population estimates. The more focused questions are, first, which of the estimators are the most consistent in reflecting the population value and, second, to what extent they under- or overestimate the population value. Because the estimators reflect different aspects of the correlation between item and score, each estimator races against itself.

5.2 Datasets used in the Study 3

For Study 3, the same real-world dataset is used as in Study 2. Now, however, the reference dataset is the original dataset of $N=4022$ test-takers ("population") from where random samples of $n=25$, 50, 100, and 200 cases are picked. A simple and straightforward statistic is calculated: the difference (d) between the sample-estimate and the population estimate. If $d=0$, the estimate was identical in the sample and in the population, if $d<0$, the population value is underestimated, and if $d>0$, the population value is overestimated.

5.3 Main results from Study 3

To outline the analysis of Study 3, Table 5 summarizes the results concerning the sources 10 and 11 of MEC. As in Studies 1 and 2, a simple ranking systemic is used. Here, if the estimator is stable in reflecting the population parameter or there is no tendency for overestimation, + 1 is given. If the estimator shows instability only in very small samples or has only insignificant overestimation, 0 is given. If the estimator shows very instable character or notable overestimation, − 1 is given.

Of the estimators in comparison, R_{REG} differs from other estimators in that it produces stable estimates with very small sample sizes also. Except R_{AC} and E_{AC}, all estimators tend to underestimate population correlation slightly, whereas R_{AC} and E_{AC} tend to give slight overestimates and E_{AC} more than R_{AC}. The impact and mechanism of the sources are discussed below.

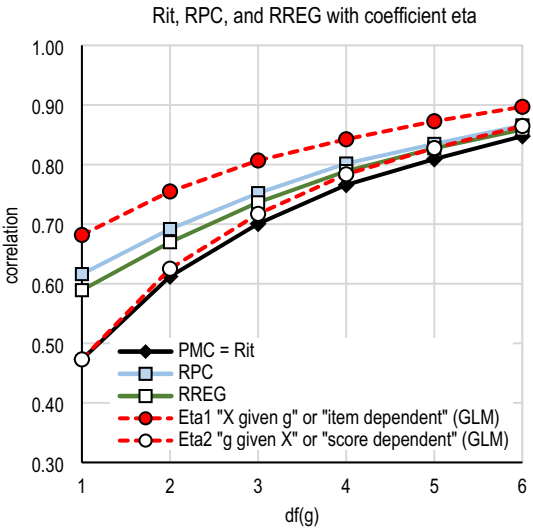

Fig. 8 Directional nature of *Rit*, R_{PC}, and R_{REG} with coefficient *eta* as a benchmark

5.4 Effect of the stability to the estimator

Obviously, all estimators produce estimates which deviate from the population value—sometimes radically (see the distribution of *RAC* as an example in Fig. 9). Especially, with small sample sizes and, specifically, if $n=25$, the estimates tend to be instable (Fig. 10). From this perspective, R_{REG} differs from the other estimators: it produces stable estimates even in the smallest sample in the study. Hence, (+1) in Table 5 for R_{REG} and (0) for the others.

5.5 Under- and overestimation in the estimates

When it comes to reflecting the population value, all estimators in comparison except R_{REG} tend to underestimate the population correlation when $df(g)<3$ (Fig. 11) which is mainly caused by instability in the estimators in the smallest sample size (cl. Fig. 10). Even if the cases with the smallest sample sizes ($n=25$) are omitted, on average, most estimators tend to underestimate population correlation mildly which may be taken as a positive signal from the viewpoint of conservativity in estimation and, hence, (+1) in Table 5 (see also Appendix 4). Notably, unlike other estimators, R_{AC} and E_{AC} tend to overestimate population correlation in an obvious manner, although the average overestimation seems to be insignificant when we compare them with the wide variety in the estimates in general (see Fig. 11). The average overestimation in R_{AC} varies 0.001–0.008 units of correlation by $df(g)$, and roughly twice in E_{AC}, that is, between 0.002 and 0.014 if the items from the smallest sample size are omitted; the overestimation is somewhat smaller if the cases from the dataset with the smallest sample size are included (see Fig. 11). Hence, in Table 5, R_{AC} is given (0) and E_{AC} (− 1). With binary items, the score for R_{AC} and E_{AC} would be + 1 as with others; the overestimation is nominal.

Table 5 Effects of sources 10 and 11 of MEC to the estimates

Source of MEC: sensitivity to	*Rit*	R_{PC}	R_{REG}	*G*	*D*	G_2	D_2	*tau-b*	R_{AC}	E_{AC}
(10) Possible instability in estimates[1]	± 0	± 0	+ 1	± 0	± 0	± 0	± 0	± 0	± 0	± 0
(11) Possible overestimation[2]	+ 1	+ 1	+ 1	+ 1	+ 1	+ 1	+ 1	+ 1	± 0	− 1
SUM	+ 1	+ 1	+ 2	+ 1	+ 1	+ 1	+ 1	+ 1	± 0	− 1

(1) scale: + 1 = stable in reflecting population parameter, ±*0* = instable only in very small samples, and –1 = very instable

(2) scale: + 1 = no tendency for overestimation, ±*0* = very small overestimation, and –1 = notable overestimation

The main reason for the instability in estimates leading also to the tendency to give under- and overestimation in the estimates is related to the dataset with the lowest sample size ($n = 25$) and binary items (see Fig. 9 above). With small sample sizes, the random selection may cause that the estimates in the sample to be far from the population—either too high (in the samples, up to + 0.40 units of correlation) or too low (as low as − 1.1 units of correlation). This phenomenon is seen, specifically, with most difficult items: the extremely difficult items in the sample appeared to be not that extreme in the population.

6 Outline: sensitivity of the estimators of correlation to MEC as a whole

Selected estimators of correlation were compared in three studies aiming to quantify to what extent they are affected by the 11 sources of MEC or relevant indicators reflecting their capability of being potential options as replacing PMC and λ_i as the weight factor w_i in MEC- and attenuated-corrected estimators of reliability. Table 6 summarizes the results. The scores would be somewhat different if only binary items were of interest.

Of the estimators in comparison, as a whole, the most vulnerable against the sources of MEC are *tau-b* (total score − 10) and PMC (− 8) which are prone to produce estimates where the magnitude of the error element related to MEC is remarkably high ($e_{wi\theta\_MEC} >> 0$), depending on, for instance, item difficulty and number of categories in the item. In this regard, the estimates by *D* (− 3) and D_2 (− 1) are notably less prone to MEC, although the magnitude of $e_{wi\theta\_MEC}$ in these estimators is also remarkable. For the latter estimators, the score would be notably higher if only binary items with extreme difficulty levels and a score with more than 20 categories were considered; *D* is not strong with polytomous items with short tests, and D_2 follows this tendency. To outline, rough quantities of MEC in these poorer behaving estimators are

$$e_{TAUi\theta\_MEC} > e_{PMCi\theta\_MEC} >> e_{Di\theta\_MEC} > e_{D_2i\theta\_MEC} >> 0. \tag{31}$$

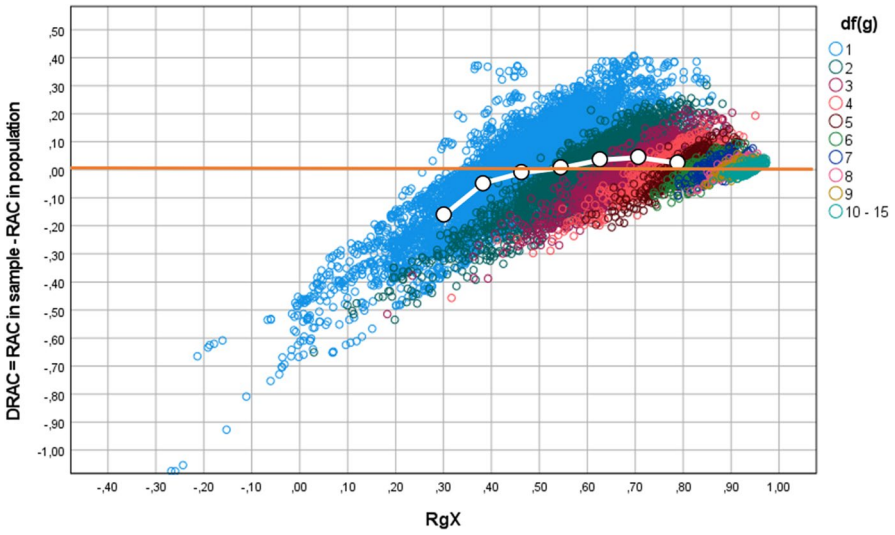

Fig. 9 Under- and overestimation in R_{AC} by *Rit* (all items, $k = 14{,}880$)

Of the better-behaving options, five are in their own class: G_2 (+17), R_{PC} (+16), R_{AC} (+16), G (+15), and E_{AC} (+15). Common to these estimators is that they do not lose information *at all* when it comes to reflecting the perfect correlation between the item and the score and, hence, in this respect, $e_{RPCi\theta\_MEC} = e_{Gi\theta\_MEC} = e_{G_2i\theta\_MEC} = e_{RACi\theta\_MEC} = e_{EACi\theta\_MEC} = 0$ in the first seven sources of MEC. As a whole, considering all 11 sources of MEC in comparison, we may conclude that these estimators bring us very near to the MEC-free condition, that is, $e_{wi\theta\_MEC} \approx 0$. Of the estimators, E_{AC} tends to produce estimates with overestimation in the real-life datasets if the item is not binary. G loses information in real-world settings with items with a wide scale and, hence, its lower score. However, with binary items, the magnitude of MEC by G is at the same level as those by R_{PC} and G_2. Also, the score of R_{REG} is very high (+14) and it could be even higher depending on how seriously the deficiencies of nominal size are penalized. The estimates by R_{REG} tend to underestimate slightly the true correlation with short tests, although the magnitude of this underestimation is insignificant.

7 Conclusion and discussion

7.1 Best options for deflation-corrected estimators of reliability

In earlier works, based on estimators of reliability in Eqs. (15) to (18), Metsämuuronen (2022a) proposed several options for MEC-corrected estimators of reliability (MCER) based on changing *Rit* and λ_i by a totally other estimators of correlation, and Metsämuuronen (2022b) has proposed some options of specific types DCERs

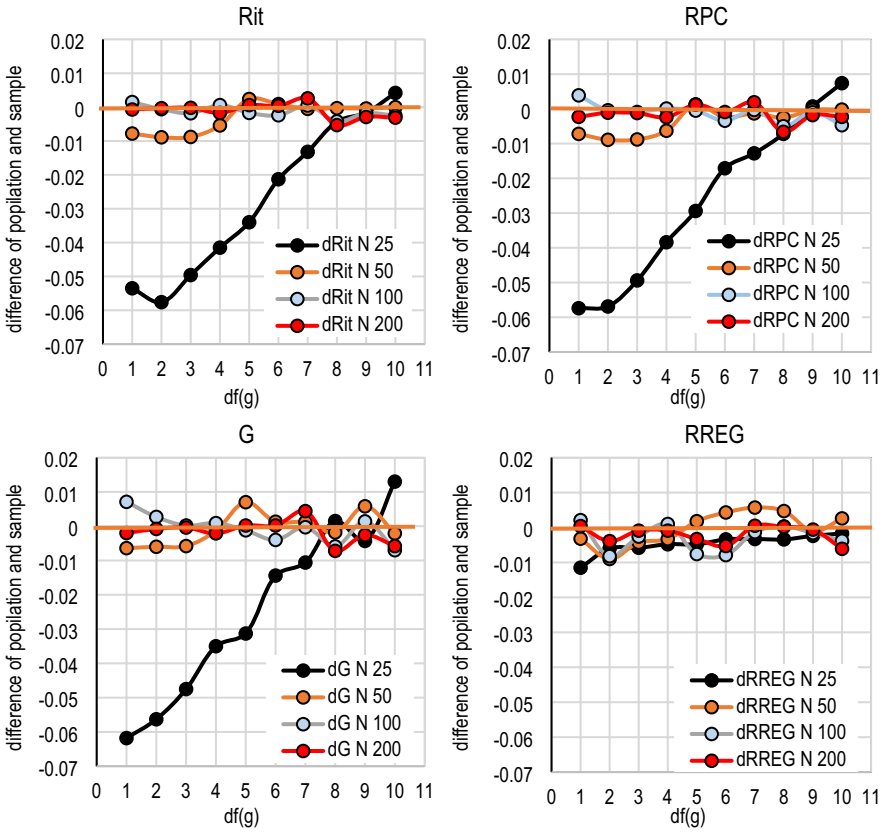

Fig. 10 Effect of sample size to the stability of the estimates

called attenuation-corrected estimators of reliability (ACER) based on attenuation-corrected *Rit* or *eta*. The task in the three sub-studies was to quantify the mechanical error in estimators of correlation and to come up with best alternatives for *Rit* and λ_i from the MEC viewpoint, that is, where $e_{wi\theta\_MEC}$ in Eqs. (7a, 7b) and (10) would be as small as possible if not totally MEC-free.

From Table 6, we may conclude that the quantities of MEC in the best-behaving estimators R_{PC}, R_{REG}, G, G_2, R_{AC}, and E_{AC} are roughly as follows:

$$e_{RREGi\theta\_MEC} > e_{RPCi\theta\_MEC} \approx e_{Gi\theta\_MEC} \approx e_{G_2i\theta\_MEC} \approx e_{RACi\theta\_MEC} \approx e_{EACi\theta\_MEC} \approx 0. \tag{32}$$

This means that, although there are small differences among the characteristics of R_{PC}, G, $G2$, R_{AC}, and E_{AC} —and R_{REG} is not far from the others—any of these estimators could be used in substituting *Rit* and λ_i in the estimators of reliability of Eqs. (15) to (18). Using these estimators, the mechanical error in reliability would be remarkably reduced, specifically, if the items are very easy, very demanding, or incrementally structured including both easy and demanding items. The last option is a typical form of a test in achievement testing. With these kinds

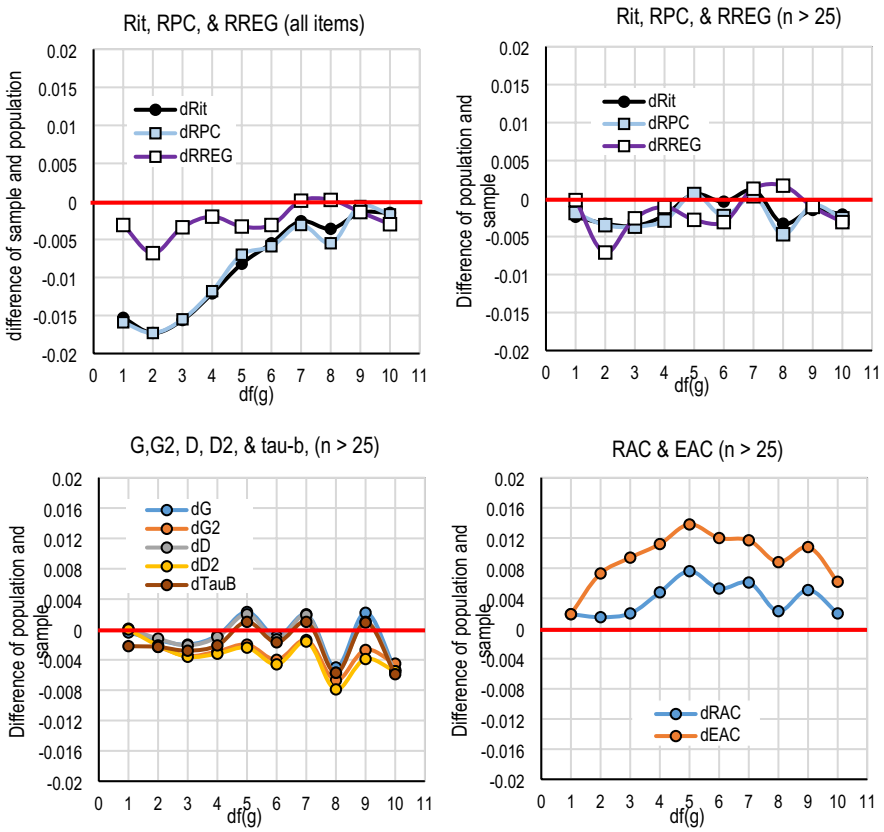

Fig. 11 Under- and overestimation in the estimates ($n > 25$, $k = 11{,}160$)

of items, *Rit* and λ_i are most vulnerable to MEC, that is, $e_{Rit\_MEC} > e_{\lambda i\_MEC} >> 0$, while the best-behaving estimators are, practically, MEC-free, that is, $e_{wi\_MEC} \approx 0$.

In practical terms, using R_{PC}, G, G_2, or R_{REG} instead of *Rit* and λ_i in the traditional estimators of reliability leads us to good options for deflation-corrected estimators of reliability and using R_{AC} or E_{AC} leads us to options for the special type of DCER, attenuation-corrected estimators of reliability. Of the latter, E_{AC} would, most probably, lead to mild overestimates. Then, as an example, with binary items, using the raw score ($\theta = X$) as a manifestation of the latent variable and $G = G_2$ as the weight factor w_i between item and score variable, based on Eq. (15), we get DCER based on alpha (ρ_{α_GiX})

$$\rho_{\alpha\_GiX} = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} \sigma_i^2 \Big/ \left(\sum_{i=1}^{k} \sigma_i G_{iX}\right)^2\right), \tag{33}$$

Table 6 Sensitivity of the estimators of correlation to MEC as a whole

Source of MEC	*Rit*	R_{PC}	R_{REG}	*G*	*D*	G_2	D_2	*tau-b*	R_{AC}	E_{AC}
(1) Discrepancy of scales[1]	− 2	+2	+2	+2	+1	+2	+1	− 2	+2	+2
(2) Item difficulty and variance[1]	− 2	+2	+2	+2	+1	+2	+1	− 2	+2	+2
(3) Distribution of the latent variable[1]	− 2	+2	+2	+2	− 2	+2	− 2	+1	+2	+2
(4) Number of categories in the item[1]	− 2	+2	+2	+2	+1	+2	+1	− 2	+2	+2
(5) Number of categories in the score[1]	− 1	+2	+1	+2	− 2	+2	− 2	− 1	+2	+2
(6) Number of items forming the score[1]	− 1	+2	+1	+2	− 2	+2	− 2	− 1	+2	+2
(7) Number of tied cases in the score[1]	− 1	+2	+1	+2	− 1	+2	− 1	− 2	+2	+2
(8) Linear or trigonometric nature [2]	+1	+1	+1	− 1	− 1	+1	+1	− 1	+1	+1
(9) Directional or symmetric nature [2]	+1	± 0	± 0	+1	+1	+1	+1	− 1	+1	+1
(10) Possible instability in estimates[3]	± 0	± 0	+1	± 0	± 0	± 0	± 0	± 0	± 0	± 0
(11) Possible overestimation[4]	+1	+1	+1	+1	+1	+1	+1	+1	± 0	− 1
SUM	− 8	+16	+14	+15	− 3	+17	− 1	− 10	+16	+15

(1) scale: + 2 = *no effect* = *MEC-free*, + 1 = *insignificant effect*, ± 0 = *unknown effect*, − 1 = *notable effect, and* − 2 = *remarkable effect lowering the estimate*

(2) scale: + 1 = trigonometric/directional nature, 0 = unknown, and − 1 = linear/symmetric nature

(3) scale: + 1 = stable in reflecting population parameter, 0 = instable only in very small samples, and − 1 = notably instable

(4) scale: + 1 = no tendency for overestimation, 0 = very small overestimation, and − 1 = notable tendency for overestimation

based on Eq. (16), we get DCER based on theta (ρ_{TH_GiX})

$$\rho_{TH\_GiX} = \frac{k}{k-1}\left(1 - 1 \Big/ \sum_{i=1}^{k} G_{iX}^2\right), \tag{34}$$

based on Eq. (17), we get DCER based on omega total (ρ_{ω_GX})

$$\rho_{\omega\_GiX} = \left(\sum_{i=1}^{k} G_{iX}\right)^2 \Big/ \left(\left(\sum_{i=1}^{k} G_{iX}\right)^2 + \sum_{i=1}^{k}\left(1 - G_{iX}^2\right)\right), \tag{35}$$

and based on Eq. (18), we get DCER based on maximal reliability (ρ_{MAX_GX})

$$\rho_{MAX\_GiX} = 1 \Big/ \left(1 + 1 \Big/ \sum_{i=1}^{k}\left(G_{iX}^2 / \left(1 - G_{iX}^2\right)\right)\right) \tag{36}$$

(see more options in Metsämuuronen 2022a). Correspondingly, R_{PC}, G_2, or R_{REG} could be used. Similarly, replacing *Rit* and λ_i by R_{AC} (or E_{AC}) leads to attenuation-corrected alpha (ρ_{α_RACiX})

$$\rho_{\alpha\_RACiX} = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k}\sigma_i^2 \Big/ \left(\sum_{i=1}^{k}\sigma_i RAC_{iX}\right)^2\right), \tag{37}$$

attenuation-corrected theta (ρ_{TH_RACiX})

$$\rho_{TH\_RACiX} = \frac{k}{k-1}\left(1 - 1 \Big/ \sum_{i=1}^{k} RAC_{iX}^2\right), \tag{38}$$

attenuation-corrected omega total (ρ_{ω_RACiX})

$$\rho_{\omega\_RACiX} = \left(\sum_{i=1}^{k} RAC_{iX}\right)^2 \Big/ \left(\left(\sum_{i=1}^{k} RAC_{iX}\right)^2 + \sum_{i=1}^{k}\left(1 - RAC_{iX}^2\right)\right), \tag{39}$$

and attenuation-corrected maximal reliability (ρ_{MAX_RACiX})

$$\rho_{MAX\_RACiX} = 1 \Big/ \left(1 + 1 \Big/ \sum_{i=1}^{k}\left(RAC_{iX}^2 \big/ \left(1 - RAC_{iX}^2\right)\right)\right) \tag{40}$$

(see Metsämuuronen 2022b). Because the estimators of correlation reflect different aspects of correlation, the estimate or reliability would vary slightly and more studies in this respect would be beneficial.

The characteristics of the estimators are not discussed here in-depth (see some initial comparisons in Metsämuuronen 2022a, b) and systematic studies in this respect would be beneficial. Obviously, using the estimators (16)–(18) outside of their original context of principal component and factor analysis is debatable. Nevertheless, that they *could* be used outside of their original contexts is consistent with a more general measurement model discussed in the article. Alternatively, DCERs based on theta, omega, and maximal reliability may be thought as an output of renewed procedures in the principal component and factor analysis where the traditional loading is replaced by an attenuation-corrected loading w_i. Studies regarding this area would be beneficial. From the viewpoint of selecting different bases for DCERs, results by Aquirre-Urreta et al. (2019) indicate that the estimators based on maximal reliability may overestimate reliability with small sample sizes. Hence, using estimators parallel to Eqs. (36) and (40) based on rho is not recommended for small sample sizes; with small sample sizes, estimators based on omega, theta, and alpha may be more useful. More studies of their behavior with finite samples would be beneficial.

Zumbo et al. (2007) and Gaderman et al. (2012) have discussed the use of alpha and theta as the bases for other types of DCERs, ordinal alpha, and ordinal theta by replacing the matrix of PMCs by a matrix of R_{PC}s in the estimation. Comparisons between ordinal alpha and theta and DCERs discussed here would be beneficial. Also, from the viewpoint of different estimators of correlation, it is good to remember Chalmers' (2017) critique against the use of R_{PC} in estimating reliability: because R_{PC} refers to theoretical and unreachable constructs, its

usefulness in assessing the accuracy of observed score may be limited. From the viewpoint of the observed score, estimators using R_{REG}, G, D, R_{AC}, and E_{AC} may be more useful. More studies in this regard would be beneficial.

7.2 Limitations of the approach

The treatment in this article has four obvious limitations. One is that the effects of the sources of MEC on the selected estimators are ranked in a rough manner to three-to-five ordinal categories. A better metric treatment of the sensitivity of estimators could bring us nearer the possibility to model and correct the possible MEC caused by selection of the estimator in the model. Second, the treatment was subjective to a certain extent and quantifying the effects with a better metric approach would advance the treatment from this viewpoint too. Third, there may be more sources of MEC not studied in the article. Anyhow, even the rough classification with 11 sources of MEC gives us a tool to assess which estimators of correlation could be "superior alternatives" when selecting the linking element in measurement modelling. An additional challenge may be that we have estimators that *over*estimate the real correlation; the article did not discuss this issue in more detail, because such estimators were not selected for the comparison (see Footnote 1 though). This issue may be worthy of more attention in future. Fourth, the practical questions of DCERs were not fully addressed—only examples were given. This area may be highly potential to study, and developing theories and practicalities related to estimation of MEC-corrected or MEC-free reliability of the test score may open new perspectives to measurement modelling settings. Systematic comparison of these kinds of new estimators would be beneficial.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41237-022-00158-y.

Acknowledgements Sincere thanks for the Counsellor of Evaluation Jukka Marjanen from FINEEC for providing the syntax for calculating bi- and polyreg correlation coefficients. Sincere thanks also for Dr. Roger Newson, research associate at the Faculty of Medicine, School of Public Health, Imperial College, London, for suggesting Greiner's relation to understand the underestimation in Somers' *D*. He also helped in getting access to other useful resources on the topic. Also, sincere thanks for Assistant Professor R. Philip Chalmers from the York University, Toronto, for pointing to the challenges of the polychoric correlation coefficient in the measurement settings in a private discussion about handling the alternative ways of estimating reliability of the test score.

Declarations

Conflict of interest There is no conflict of interest.

References

Anselmi P, Colledai D, Robusto E (2019) A comparison of classical and modern measures of internal consistency. Front Psychol 10:2714. https://doi.org/10.3389/fpsyg.2019.02714

Aquirre-Urreta M, Rönkkö M, McIntosh CN (2019) A cautionary note on the finite sample behavior of maximal reliability. Psychol Methods 24(2):236–252. https://doi.org/10.1037/met0000176

Armor D (1973) Theta reliability and factor scaling. Sociol Methodol 5:17–50. https://doi.org/10.2307/270831

Berry KJ, Johnston JE, Mielke PW Jr (2018) The measurement of correlation. A permutation statistical approach. Springer, Cham. https://doi.org/10.1007/978-3-319-98926-6

Byrne BM (2016) Structural equation modelling with AMOS Basic concepts, applications, and programming. Third Edition. Routledge

Chalmers RP (2017) On misconceptions and the limited usefulness of ordinal alpha. Educ Psychol Measur 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: Lance CE, Vanderberg RJ (eds) Statistical and methodological myths and urban legends. Routledge, Milton Park, pp 309–326. 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. Educ Psychol Measur 72(1):52–67. https://doi.org/10.1177/0013164411407315

Clemans WV (1958) An index of item-criterion relationship. Educ Psychol Measur 18(1):167–172. https://doi.org/10.1177/001316445801800118

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

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

Drasgow F (1986) Polychoric and polyserial correlations. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 7. Wiley, Hoboken, pp 68–74

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

Gadermann AM, Guhn M, Zumbo BD (2012) Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. Pract Assess Res Eval 17(3):1–13. https://doi.org/10.7275/n560-j767

Göktaş A, İşçi OA (2011) Comparison of the most commonly used measures of correlation for doubly ordered square contingency tables via simulation. Metodološki zvezki [Methodological Notebooks] 8(1):17–37. https://www.stat-d.si/mz/mz8.1/goktas.pdf.

Gonzalez R, Nelson TO (1996) Measuring ordinal correlation in situations that contain tied scores. Psychol Bull 119(1):159–165. https://doi.org/10.1037/0033-2909.119.1.159

Goodman LA, Kruskal WH (1954) Measures of correlation for cross classifications. J Am Stat Correl 49(268):732–764. https://doi.org/10.1080/01621459.1954.10501231

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

Green SB, Yang Y (2015) Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. Educ Meas Issues Pract 34(4):14–20. https://doi.org/10.1111/emip.12100

Greene VL, Carmines EG (1980) Assessing the reliability of linear composites. Sociol Methodol 11:160–217. https://doi.org/10.2307/270862

Greiner R (1909) Über das Fehlersystem der Kollektivmaßlehre [Of the error systemic of collectives]. J Math Phys (Zeitschift fur Mathematik und Physik) 57:121–158, 225–260, 337–373

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

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

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

Higham PA, Higham DP (2019) New improved gamma: enhancing the accuracy of Goodman-Kruskal's gamma using ROC curves. Behav Res Methods 51(1):108–125. https://doi.org/10.3758/s13428-018-1125-5

IBM (2017) IBM SPSS Statistics 25 Algorithms. IBM. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/25.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf
Jackson RWB, Ferguson GA (1941) Studies on the reliability of tests. Department of Educational Research, University of Toronto
Kaiser HF, Caffrey J (1965) Alpha factor analysis. Psychometrika 30:1–14. https://doi.org/10.1007/BF02289743
Kendall M (1938) A new measure of rank correlation. Biometrika 30(1/2):81–93. https://doi.org/10.2307/2332226
Kendall MG (1948) Rank correlation methods, 1st edn. Charles Griffin and Co Ltd
Kendall M (1949) Rank and product–moment correlation. Biometrika 36(1/2):177–193. https://doi.org/10.2307/2332540
Kim J-O, Mueller CW (1978) Introduction to Factor Analysis: What It Is and How to Do It. Series: Quantitative Applications in the Social Sciences, n.o. 13. Sage Publication, Inc.
Kuder GF, Richardson MW (1937) The theory of the estimation of test reliability. Psychometrika 2(3):151–160. https://doi.org/10.1007/BF02288391
Kvålseth TO (2017) An alternative measure of ordinal correlation as a value-validity correction of the Goodman-Kruskal gamma. Commun Stat—Theory Methods 46(21):10582–10593. https://doi.org/10.1080/03610926.2016.1239114
Lancaster HO, Hamdan MA (1964) Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters. Psychometrika 29:383–391. https://doi.org/10.1007/BF02289604
Lavrakas PJ (2008) Attenuation. In: Lavrakas PJ (ed) Encyclopedia of survey methods. Sage Publications, Inc., Thousand Oaks. https://doi.org/10.4135/9781412963947.n24
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 DB (1996) Reliability of measurement in psychology: from Spearman-Brown to maximal reliability. Psychol Methods 1:97–108. https://doi.org/10.1037/1082-989X.1.1.98
Livingston SA, Dorans NJ (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 FM (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 FM, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley Publishing Company, Boston
Lorenzo-Seva U, Ferrando PJ (2015) POLYMAT-C: a comprehensive SPSS program for computing the polychoric correlation matrix. Behav Res Methods 47:884–889. https://doi.org/10.3758/s13428-014-0511-x
Martin WS (1973) The effects of scaling on the correlation coefficient: a test of validity. J Mark Res 10(3):316–318. https://doi.org/10.2307/3149702
Martin WS (1978) Effects of scaling on the correlation coefficient: additional considerations. J Mark Res 15(2):304–308. https://doi.org/10.1177/002224377801500219
Martinson EO, Hamdan MA (1972) Maximum likelihood and some other asymptotical efficient estimators of correlation in two-way contingency tables. J Stat Comput Simul 1(1):45–54. https://doi.org/10.1080/00949657208810003
McDonald RP (1970) Theoretical canonical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. Br J Math Stat Psychol 23:1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x
McDonald RP (1985) Factor analysis and related methods. Lawrence Erlbaum Associates, Mahwah
McDonald RP (1999) Test theory: a unified treatment. Lawrence Erlbaum Associates, Mahwah
McNeish D (2017) Thanks coefficient alpha, we'll take it from here. Psychol Methods 23(3):412–433. https://doi.org/10.1037/met0000144
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
Meade AW (2010) Restriction of range. In: Salkind NJ (ed) Encyclopedia of research design. SAGE Publications Inc, Thousand Oaks, pp 1278–1280. https://doi.org/10.4135/9781412961288.n309
Mendoza JL, Mumford M (1987) Corrections for attenuation and range restriction on the predictor. J Educ Stat 12(3):282–293. https://doi.org/10.3102/10769986012003282
Metsämuuronen J (2016) Item–total correlation as the cause for the underestimation of the alpha estimate for the reliability of the scale. GJRA - Global J Res Anal 5(1):471–477. https://www.world

widejournals.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, Inc, Thousand Oaks
Metsämuuronen J (2020a) Somers' D as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. Int J Educ Methodol 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. Int J Educ Methodol 6(2):297–317. https://doi.org/10.12973/ijem.6.2.297
Metsämuuronen J (2020c) Directional nature of the product–moment correlation coefficient. Preprint at https://doi.org/10.13140/RG.2.2.36815.71843
Metsämuuronen J (2021a) Goodman-Kruskal gamma and dimension-corrected gamma in educational measurement settings. Int J Educ Methodol 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. https://doi.org/10.1007/s41237-021-00138-8
Metsämuuronen J (2022a) Deflation-corrected estimators of reliability. Front Psychol 12:748672. https://doi.org/10.3389/fpsyg.2021.748672
Metsämuuronen J (2022b) Attenuation-corrected reliability and some other MEC-corrected estimators of reliability. Preprint at https://doi.org/10.13140/RG.2.2.22647.75689/1
Metsämuuronen J (2022c) Mechanical attenuation in eta squared and some related consequences. Attenuation-corrected eta and eta squared, negative values of eta, and their relation to Pearson correlation. Preprint at https://doi.org/10.13140/RG.2.2.29569.58723
Moses T (2017) A review of developments and applications in item analysis. In: Bennett R, von Davier M (eds) Advancing human assessment. The methodological, psychological and policy contributions of ETS. Educational Testing Service. Springer Open, Cham, pp 19–46. https://doi.org/10.1007/978-3-319-58689-2_2
Newson R (2002) Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and Median Differences. Stata J 2(1):45–64. http://www.stata-journal.com/sjpdf.html?articlenum=st0007
Newson R (2008) Identity of Somers' *D* and the rank biserial correlation coefficient. http://www.rogernewsonresources.org.uk/miscdocs/ranksum1.pdf
Novick MR, Lewis C (1967) Coefficient alpha and the reliability of composite measurements. Psychometrika 32(1):1–13. https://doi.org/10.1007/BF02289400
Olsson (1979) Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika 44:443–460. https://doi.org/10.1007/BF02296207
Olsson U (1980) Measuring correlation in ordered two-way contingency tables. J Mark Res 17(3):391–394. https://doi.org/10.1177/002224378001700315
Olsson U, Drasgow F, Dorans NJ (1982) The polyserial correlation coefficient. Psychometrika 47:337–347. https://doi.org/10.1007/BF02294164
Oosterhof AC (1976) Similarity of various item discrimination indices. J Educ Meas 13(2):145–150. https://doi.org/10.1111/j.1745-3984.1976.tb00005.x
Pearson K (1896) Mathematical contributions to the theory of evolution III. regression, heredity, and panmixia. Philos Trans R Soc Lond. Ser a, Contain Papers Math Phys Character 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. Philos Trans R Soc A Math, Phys Eng Sci 195(262–273):1–47. https://doi.org/10.1098/rsta.1900.0022
Pearson K (1903) I. Mathematical contributions to the theory of evolution.—XI. On the influence of natural selection on the variability and correlation of organs. Philos Trans R Soc A Math Phys Eng Sci 200(321–330):1–66. https://doi.org/10.1098/rsta.1903.0001
Pearson K (1905) On the general theory of skew correlation and non-linear regression. Dulau and Co., London. https://archive.org/details/ongeneraltheory00peargoog/page/n3
Pearson K (1909) On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. Biometrika 7(1–2):96–105. https://doi.org/10.1093/biomet/7.1-2.96
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 modelling approach. Br J Math Stat Psychol 57(1):21–27. https://doi.org/10.1348/000711004849295

Sackett PR, Yang H (2000) Correction for range restriction: an expanded typology. J Appl Psychol 85(1):112–118. https://doi.org/10.1037/0021-9010.85.1.112

Sackett PR, Lievens F, Berry CM, Landers RN (2007) A cautionary note on the effect of range restriction on predictor intercorrelations. J Appl Psychol 92(2):538–544. https://doi.org/10.1037/0021-9010.92.2.538

Schmidt FL, Shaffer JA, 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. Pers Psychol 61(4):827–868. https://doi.org/10.1111/j.1744-6570.2008.00132.x

Sheskin DJ (2011) Handbook of parametric and nonparametric statistical procedures, 5th edn. Chapman and Hall/CRC, London

Siegel S, Castellan NJ Jr (1988) Nonparametric statistics for the behavioral sciences, 2nd edn. McGraw-Hill, New York

Sirkin MR (2006) Statistics of the social science, 3rd edn. SAGE Publications, Inc, Thousand Oaks

Somers RH (1962) A new asymmetric measure of correlation for ordinal variables. Am Sociol Rev 27(6):799–811. https://doi.org/10.2307/2090408

Spearman C (1904) The proof and measurement of correlation between two things. Am J Psychol 15(1):72–101. https://doi.org/10.2307/1412159

Trizano-Hermosilla I, Alvarado JM (2016) Best alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. Front Psychol 7:769. https://doi.org/10.3389/fpsyg.2016.00769

Walk MJ, Rupp AA (2010) Pearson product-moment correlation coefficient. In: Salkind NJ (ed) Encyclopedia of research design. SAGE Publications, Inc., Thousand Oaks, pp 1022–1026. https://doi.org/10.4135/9781412961288.n309

Wholey JS, Hatry HP, Newcomer KE (eds) (2015) Handbook of practical program evaluation, 4th edn. Jossey-Bass, San Francisco

Yang H (2010) Factor loadings. In: Salkind NJ (ed) Encyclopedia of research design. SAGE Publications, Thousand Oaks, pp 480–483. https://doi.org/10.4135/9781412961288.n309

Zaionts C (2021) Real Statics Using Excel. Polychoric Correlation using Solver. http://www.real-statistics.com/correlation/polychoric-correlation/polychoric-correlation-using-solver/

Zumbo BD, Gadermann AM, Zeisser C (2007) Ordinal versions of coefficients alpha and theta for Likert rating scales. J Modern Appl Stat Methods 6(1):21–29. https://doi.org/10.22237/jmasm/1177992180