Appendix 1. Conceptual discussion of deflation in the measurement model

Traditional measurement model without attenuation and deflation

Consider a measurement model based on latent variable modelling assuming a latent trait or a latent variable \( \theta \) manifested as the score variable \( (X) \) and driving the observed responses in \( (x_i) \) in a test item \( (g_i) \) (see, McDonalds, 1985, 1999; Raykov & Marcoulides, 2010, 2013):

\[
x_i = \lambda_i \theta + e_i,
\]

where \( \lambda_i \) is the weighting element between \( \theta \) and \( x_i \) \((-1 \leq \lambda_i \leq +1) \) and \( e_i \) is the item-wise measurement error. For convenience, \( x_i \) and \( \theta \) are assumed standardized and centered so that \( x_i, \theta \sim N(0,1) \). Then, after taking variances on both sides,

\[
VAR(x_i) = VAR(\lambda_i \theta) + VAR(e_i),
\]

we get the item-wise error variance

\[
\psi_i^2 = 1 - \lambda_i^2.
\]

Then, the classic relation of \( T, X \) and \( E \) can be written as the sum of the test items as

\[
X = T + E = \sum_{i=1}^{k} x_i = \sum_{i=1}^{k} \lambda_i \theta + \sum_{i=1}^{k} e_i
\]

from which the error variance of the test score can be written as

\[
\sum_{i=1}^{k} \psi_i^2 = \sum_{i=1}^{k} (1 - \lambda_i^2).
\]

Usually in these conceptualizations, \( \theta \) is operationalized as a factor score variable \( (F) \) or principal component score \( (P) \), that is, a linear combination of weighted items, and \( \lambda_i \) as a factor loading. In a graphical form, the traditional (general congeneric) measurement model can be expressed as in Figure 1.

![Figure 1. A congeneric one-factor measurement model](image)

This operationalization is strictly used in such estimator of reliability as theta, omega total, and maximal reliability.
Element of MEC in the measurement model

The traditional measurement model implies that the linking coefficient $\lambda_i$ is errorless. However, as being (essentially) correlations between the items and the score, the loadings are negatively biased by attenuation and deflation and, hence, the observed estimates by these estimators are underestimates. If we include the element of mechanical error in the estimators of correlation (MEC) in the model, it obviously changes the model to some extent.

Let us consider a more general model parallel to Eq (1) where $\theta$ may be manifested as a varying type of relevantly formed compilation of items including a raw score ($\theta_X$), principal component score variable ($\theta_{PC}$), factor score variable ($\theta_{FA}$), or a theta score formed by the item response theory (IRT) or Rasch modelling ($\theta_{IRT}$). Also, instead of exclusively a principal component- or factor loading, each item gets a general unique weight $w_i$ that links $\theta$ with $g_i$. It is reasonable to assume that the general weight element $w_i$ is a coefficient of correlation ($-1 \leq w_i \leq +1$) including principal component loading, factor loading, as well as a modification of discriminating parameter $a$ in IRT modelling. Then, a measurement model parallel to in Eq. (1) is

$$x_i = w_i \theta + e_i$$  \hspace{1cm} (6)

In a graphical form, this model looks, essentially, the same as the factor model (Fig. 2) although the symbols are more general.

Figure 2. A more general one-latent variable measurement model

The consequence of attenuation and deflation caused by MEC is that the magnitude of the observed or MEC-defected correlation ($w_{i,\text{Observed}}$) between $\theta$ and $g_i$ tends to be lower than the MEC-corrected (MECC) correlation ($w_{i,\text{MECC}}$) and MEC-free (MECF) correlation. Then, when obtaining MEC,

$$w_{i,\text{MECC}} = w_{i,\text{Observed}} + e_{w\theta,\text{MEC}}$$  \hspace{1cm} (7)

where $e_{w\theta,\text{MEC}}$ refers to the systematical mechanical error related to the specific weight factor $w$, specific item $i$, and a specific manifestation of the latent variable ($\theta$). When MEC is obtained, $e_{w\theta,\text{MEC}} > 0$ although its magnitude is generally unknown, and it depends on several factors.

If we select wisely the weight factor, the magnitude of error component related to attenuation and deflation is zero when the MEC-free condition is met and near zero when MEC-corrected condition is met, that is, $e_{w\theta,\text{MEC}} \approx 0$. In a good case, the MEC-corrected condition would be as near the MEC-free condition as possible, that is,

$$x_i = w_{i,\text{MECC}} \times \theta + \left(e_{i,\text{Random}} + e_{w\theta,\text{MEC}}\right)$$  \hspace{1cm} (8)
As in Eq. (2), we take variances of both sides and organize the terms. We recall that all generally used estimators of correlation give identical estimate of the correlation for original variables \((g_i \text{ and } \theta)\) and for the standardized versions of the variables \((\text{STD}(g_i) \text{ and } \text{STD}(\theta))\). Therefore, without loss of generality, irrespective of their factual form, we can assume that \(x_i, \theta \sim N(0,1)\) and then \(e_{i,\text{MECC}} \sim N(0, \psi_{i,\text{MECC}}^2)\), where
\[
\psi_{i,\text{MECC}}^2 = 1 - w_{i,\text{MECC}}^2,
\]
where \(\psi_{i,\text{MECC}}^2\) refers to the MEC-corrected error variance related to a single item and specific weighting element. Then, parallel to Eq. (4), the MEC-corrected model can be written as
\[
\sum_{i=1}^{l} x_i = \sum_{i=1}^{l} (w_{i,\text{MECC}} \times \theta) + \sum_{i=1}^{k} e_{i,\text{Random}}.
\]
In a graphical form, the deflation- or MEC-corrected measurement model looks as in Figure 3.

![Deflation-corrected one-latent variable measurement model](image)

**Figure 3.** Deflation-corrected one-latent variable measurement model
Appendix 2. Numerical examples of deflation

Case 1: A test of extreme difficulty level

Assume a hypothetical dataset of five items with 0 to 2 points (Table 1a). This could be a screening test of understanding instructions where all native speakers are expected to get full marks while second language speakers or those with some learning difficulty may make some mistakes in the test items. Additionally, Table 1a includes item arrays leading to maximal covariance and maximal correlation. With two variables ($g_2$ and $g_4$), the observed correlation reaches the maximal value; hence $R_{AC} = 1$ as also indicated by $R_{PC}$ and $G$.

Table 1a. Hypothetical dataset with extremely easy items

| ID | g1 | g2 | g3 | g4 | g5 | Score (X) | g1 | g2 | g3 | g4 | g5 |
|----|----|----|----|----|----|-----------|----|----|----|----|----|
| 1  | 2  | 0  | 2  | 0  | 2  | 6         | 1  | 0  | 0  | 0  | 0  |
| 2  | 2  | 2  | 1  | 2  | 0  | 7         | 2  | 2  | 1  | 0  | 1  |
| 3  | 1  | 2  | 0  | 2  | 0  | 7         | 2  | 2  | 2  | 2  | 1  |
| 4  | 2  | 2  | 0  | 2  | 2  | 8         | 2  | 2  | 2  | 2  | 2  |
| 5  | 2  | 2  | 2  | 2  | 1  | 9         | 2  | 2  | 2  | 2  | 1  |
| 6  | 2  | 2  | 1  | 2  | 1  | 9         | 2  | 2  | 2  | 2  | 1  |
| 7  | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 8  | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 9  | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 10 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 11 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 12 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 13 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 14 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 15 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 16 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 17 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 18 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 19 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |
| 20 | 2  | 2  | 2  | 2  | 2  | 10        | 2  | 2  | 2  | 2  | 2  |

$p = 0.975$ 0.967 0.925 0.900 0.900 0.975 0.967 0.925 0.900 0.900
$p = 0.975$ 0.967 0.925 0.900 0.900

$\sigma^2_i = 0.048$ 0.190 0.228 0.360 0.260 $\sigma^2_i = 0.048$ 0.190 0.228 0.360 0.260

$\rho_{X_{ob}} = 0.429$ 0.616 0.418 0.760 0.415 $\rho_{X_{ob}} = 0.429$ 0.616 0.418 0.760 0.415

$R_{AC} = 0.697$ 1 0.551 1 0.464 $R_{AC} = 0.697$ 1 0.551 1 0.464

$R_{AC} = \rho_{X_{ob}} / \rho_{X_{ob}}$
Table 1b. Statistics based on Table 1a

| item | $\sigma_i^2$ | $R_{PC}$ | $G$ | $D$ | $G_2$ | $D_2$ | $\sigma \times \rho_{\phi X}$ | $\sigma \times R_{AC}$ | $\sigma \times X$ | $\sigma \times D$ | $\sigma \times G$ | $\sigma \times G_2$ | $\sigma \times D_2$ |
|------|-------------|---------|-----|-----|-------|------|----------------------|-------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| g1   | 0.048       | 0.743   | 0.889 | 0.842 | 0.842 | 0.094 | 0.152                | 0.162            | 0.194       | 0.194       | 0.194       | 0.194       | 0.194       |
| g2   | 0.190       | 1      | 1    | 1    | 1     | 0.269 | 0.436                | 0.436            | 0.436       | 0.436       | 0.436       | 0.436       | 0.436       |
| g3   | 0.228       | 0.673   | 0.778 | 0.757 | 0.800 | 0.800 | 0.199                | 0.263            | 0.321       | 0.371       | 0.361       | 0.372       | 0.372       |
| g4   | 0.360       | 1      | 1    | 0.972 | 0.972 | 0.456 | 0.600                | 0.600            | 0.600       | 0.583       | 0.583       | 0.583       | 0.583       |
| g5   | 0.260       | 0.693   | 0.731 | 0.717 | 0.756 | 0.742 | 0.212                | 0.237            | 0.354       | 0.373       | 0.366       | 0.379       | 0.379       |
| SUM  | 1.085       |         |      |      |       |       |                      |                  |             |             |             |             |             |

Table 1c. Estimates of ACERs and MCERs of the score based on Table 1a

| Eq. | 1 | 2 | 3 | 11 | 12 | 13 | 5 | 6 | 7 | 8 | 9 |
|-----|---|---|---|----|----|----|---|---|---|---|---|---|
|     | $\rho_{a, R_X}$ | $\rho_{TH}$ | $\rho_a$ | $\rho_{a, RACX}$ | $\rho_{TH, RACX}$ | $\rho_{a, RACX}$ | $\rho_{a, RPCX}$ | $\rho_{a, GX}$ | $\rho_{a, DX}$ | $\rho_{a, G_{0X}}$ | $\rho_{a, D_{0X}}$ |
|     | 0.352 | 0.631 | - | 0.774 | 0.834 | 0.873 | 0.863 | 0.902 | 0.886 | 0.910 | 0.894 |

Statistics for estimating reliability of the score are collected in Table 1b. Estimates by traditional estimators as well as by ACERs and MCERs are collected in Table 1c. In Table 1a, estimates by $G$ and $D$ are standard outputs of a statistical software package although they are easy to calculate manually. Here, $D$ refers to $D(g|X)$, that is, “$g$ given $X$” which is usually worded as “$X$ dependent” in the software packages (see the discussion of naming and the proper direction in, e.g., Metsämuuronen, 2020a, 2021a, 2021b). $G_2$ and $D_2$ developed to correct the obvious underestimation obtained by $G$ and $D$ when the number of categories exceeded four ($G$) or three ($D$) are calculated by the procedure suggested in Metsämuuronen (2021a):

$$ G_2 = G \times \left(1 + (1 - \text{abs}(G)) \times A \right) $$

where $A = \frac{df(g) - 1}{df(g)} \left(1 - \frac{1}{df(g)} \right)^2$ and $df(g) = \text{(number of categories–1)}$. As an example, for item $g_3$ with 3 categories, $A = \left((2-1)/2\right) \times (1-1/2)^2 = 0.125$. Then, $G_2 = 0.778 \times (1 + (1-0.778) \times 0.125) = 0.800$ and $D_2 = 0.757 \times (1 + (1-0.757) \times 0.125) = 0.780$. $R_{PC}$ is estimated manually by enhancing Zaiontz’s (2022) simplified version of the two-step procedure by Martinson and Hamdan (1970).

Four points highlighted from Case 1 are relevant also in Case 2. First, based on Tables 1a and 1b, using Eq. (1), the traditional coefficient alpha underestimates the reliability in an obvious manner: $\hat{\alpha} = 5/4 \times (1-1.085/1.229^2) = 0.352$. The low value is caused by reduction in item variances leading to MEC in observed $\rho_{\phi X}$; even at the highest, given the dataset, $\rho_{\phi X}$ can reach values $\rho_{\phi X, \text{Max}} = 0.616 - 0.894$. Eq. (11) gives an attenuation-corrected estimate $\hat{\alpha}_{RAC} = 5/4 \times (1-1.085/1.687^2) = 0.774$. Although the correction of attenuation or deflation in reliability based on the alpha formula is remarkable (0.42 units of reliability), it seems conservative in comparison with the more advanced ACERs: Eq. (12) gives an estimate $\hat{\rho}_{TH, RACX} = 5/4 \times (1-1/3.004) = 0.834$ and Eq. (13) $\hat{\rho}_{\phi, RACX} = 3.712^2 / (3.712^2 + 1.996) = 0.873$; all are notably higher than the traditional alpha and theta ($\hat{\rho}_{TH\theta} = 0.631$). In comparison with the different ACERs, the original alpha seems deflated by 0.48–0.52 units of reliability. Notably, estimates by omega and maximal reliability cannot be calculated because the correlation matrix is not positively definite.
Second, because some of the MCERs based on changing the entire coefficient have concrete interpretations, their estimates may be valuable benchmarks to the deflation in the traditional alpha. The estimator based on $R_{PC}$ (Eq. 5) gives the estimate $\hat{\alpha}_{RPC} = 5/4 \times (1 - 1.085/1.872^2) = 0.863$, the estimator based on $G$ (Eq. 6) $\hat{\alpha}_G = 5/4 \times (1 - 1.085/1.973^2) = 0.902$, and the estimator based on $D$ (Eq. 8) gives the value $\hat{\alpha}_D = 5/4 \times (1 - 1.085/1.929^2) = 0.886$. The estimates by $G$ and $D$ strictly indicates the proportion of logically ordered test takers in the item after they are ordered by the score; this proportion can be calculated by $p = 0.5 \times G + 0.5$ and $p = 0.5 \times D(g|X) + 0.5$ derived from Metsämuuronen (2021b). For example, by using $D$, in item $g_1$ this proportion is $p = 0.5 \times 0.842 + 0.5 = 0.921$, that is, 92.1% of all observations in item $g_1$ are in a logical order after ordered by the score. In all items together, the average proportion is 92.9%. Hence, the set of items can discriminate very effectively those who got full marks from other test takers. Thus, it seems that attenuation-corrected values reflect more accurately the MEC-free reliability than the original estimate by alpha.

Third, that the magnitude of the estimates by $\alpha_{RAC}$ is lower than the one by $\alpha_{RPC}$ is not a general characteristic. In Case 2, it is to be seen that $\hat{\alpha}_{RAC} > \hat{\alpha}_{RPC}$. That the estimate based on $G$ is higher than the one by $R_{PC}$ is not a general characteristic either; it is also a coincidence in the dataset. In real-life settings with two or three categories in the item, $G$ and $R_{PC}$ seem to give estimates that are quite close to each other (see simulations in Metsämuuronen, 2021a, 2021d). However, that the magnitude of the estimates by $D$ are lower than those by $G$ is expected because, with the same pairs of variables, $D$ tends to give more conservative estimates of association than $G$ (see Metsämuuronen, 2021b). Also, that the magnitude of the estimates is higher when using $G_2$ and $D_2$ in comparison with $G$ and $D$ is expected because $G_2$ and $D_2$ are developed to correct the obvious underestimation obtained by $G$ and $D$ when the number of categories exceeded three ($D$) or four ($G$). Using Eq. (7) based on $G_2$, we get an estimate $\hat{\alpha}_{G_2} = 5/4 \times (1 - 1.085/1.996^2) = 0.910$ and by Eq. (9) based on $D_2$, we get $\hat{\alpha}_{D_2} = 5/4 \times (1 - 1.085/1.953^2) = 0.894$.

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

**Case 2: Incrementally structured dataset**

Assume a hypothetical dataset as in Table 2a of five items with 0 to 2 points with incremental difficulty level of items. This could be a short-ish sub-test of “Algebra” as a part of a longer achievement test of mathematics. The statistics for estimating reliability of the score in Table 2a are collected in Table 2b. As in Table 1b, the estimates by $G$ and $D = D(g|X)$, are standard output of a statistical software package, $G_2$, and $D_2$ are calculated manually as above, and $R_{PC}$ is estimated.
manually by enhancing Zaiontz’s procedure (2022). Different estimates of ACERs and MCERs are collected in Table 2c.

Basically, the main result is the same as in Case 1: reliability estimated by the traditional coefficient alpha (\( \hat{\alpha} = 0.411 \)), theta (\( \hat{\rho}_{TH} = 0.531 \)), and omega (\( \hat{\rho}_{\omega} = 0.563 \)) are deflated because the test comprises both very easy and difficult items—even at the highest in the given dataset, PMC in the extreme items could not exceed values \( \rho_{\xi_{\text{Max}}} = 0.451–0.482 \). The estimates of the ACERs by Eqs. (10), (11), and (12) are \( \hat{\alpha}_{RAC} = 0.790 \), \( \hat{\rho}_{TH,RAC} = 0.838 \), \( \hat{\rho}_{\omega,RAC} = 0.881 \), respectively. The estimate by the ACER based on the alpha formula comes quite close to the ones by other types of MEC-corrected estimates by Eqs. (5), (6), and (7) (\( \hat{\alpha}_{RPC} = 0.787 \), \( \hat{\alpha}_{G} = 0.785 \), and \( \hat{\alpha}_{\omega} = 0.806 \), respectively). In the case, the differences between \( \alpha_{RAC} \) and other deflation-corrected estimates based on alpha are subtle but the difference between these and the traditional \( \alpha \) is notable; the traditional alpha seems to underestimate reliability around 0.38–0.47 units of reliability, traditional theta seems to underestimate reliability around 0.25 units, and traditional omega around 0.24 units of correlation.

### Table 2a. Hypothetic dataset with extremely easy items

| ID | g1 | g2 | g3 | g4 | g5 | Score (X) | g1 | g2 | g3 | g4 | g5 |
|----|----|----|----|----|----|-----------|----|----|----|----|----|
| 1  | 1  | 1  | 0  | 0  | 0  | 2         | 0  | 0  | 0  | 0  | 0  |
| 2  | 2  | 0  | 0  | 0  | 0  | 2         | 1  | 0  | 0  | 0  | 0  |
| 3  | 2  | 0  | 0  | 0  | 0  | 2         | 2  | 0  | 0  | 0  | 0  |
| 4  | 0  | 1  | 2  | 0  | 0  | 3         | 2  | 0  | 0  | 0  | 0  |
| 5  | 2  | 0  | 2  | 0  | 0  | 4         | 2  | 1  | 0  | 0  | 0  |
| 6  | 2  | 2  | 0  | 0  | 0  | 4         | 2  | 1  | 0  | 0  | 0  |
| 7  | 2  | 2  | 0  | 0  | 0  | 4         | 2  | 2  | 0  | 0  | 0  |
| 8  | 2  | 2  | 0  | 0  | 0  | 4         | 2  | 2  | 0  | 0  | 0  |
| 9  | 2  | 2  | 0  | 0  | 0  | 4         | 2  | 2  | 0  | 0  | 0  |
| 10 | 2  | 2  | 0  | 1  | 0  | 5         | 2  | 2  | 1  | 0  | 0  |
| 11 | 2  | 2  | 0  | 0  | 1  | 5         | 2  | 2  | 1  | 0  | 0  |
| 12 | 2  | 2  | 1  | 0  | 0  | 5         | 2  | 2  | 2  | 0  | 0  |
| 13 | 2  | 2  | 2  | 0  | 0  | 6         | 2  | 2  | 2  | 0  | 0  |
| 14 | 2  | 2  | 2  | 0  | 0  | 6         | 2  | 2  | 2  | 0  | 0  |
| 15 | 2  | 2  | 1  | 2  | 0  | 7         | 2  | 2  | 2  | 1  | 0  |
| 16 | 2  | 2  | 2  | 1  | 0  | 7         | 2  | 2  | 2  | 1  | 0  |
| 17 | 2  | 2  | 2  | 2  | 0  | 8         | 2  | 2  | 2  | 2  | 0  |
| 18 | 2  | 2  | 0  | 2  | 2  | 8         | 2  | 2  | 2  | 2  | 0  |
| 19 | 2  | 2  | 2  | 2  | 0  | 8         | 2  | 2  | 2  | 2  | 1  |
| 20 | 2  | 2  | 2  | 2  | 0  | 8         | 2  | 2  | 2  | 2  | 2  |

| \( p \) | 0.925 | 0.750 | 0.500 | 0.300 | 0.075 | \( p \) | 0.925 | 0.750 | 0.500 | 0.300 | 0.075 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| \( \sigma_i^2 \) | 0.228 | 0.650 | 0.900 | 0.740 | 0.228 | \( \sigma_i^2 \) | 0.2275 | 0.650 | 0.900 | 0.740 | 0.228 |
| \( \rho_{\xi_{\text{Obs}}} \) | 0.378 | 0.675 | 0.469 | 0.828 | 0.295 | \( \rho_{\xi_{\text{Max}}} \) | 0.482 | 0.767 | 0.886 | 0.828 | 0.451 |
| \( R_{AC} \) | 0.785 | 0.880 | 0.529 | 0.991 | 0.655 | \( R_{AC} = \rho_{\xi_{\text{Obs}}} / \rho_{\xi_{\text{Max}}} \) |
The results are discussed in the main text.

Case 3. Larger simulation of the behavior of $R_{AC}$ and attenuation-corrected estimator of reliability

A real-life dataset of 4,022 nationally representative test-takers of a mathematics test with 30 binary items (FINEEC, 2018) is used as the “population” in simulation of the behavior of $R_{AC}$ and ACERs in the real-life testing settings. In the original dataset, the difficulty levels of the items range $0.24 < \rho < 0.95$, item–score correlation range 0.332 < $R_{it} < 0.627$, and the traditional estimates of reliability are $\rho_a = 0.885$, $R_{TH} = 0.890$, $\omega_a = 0.887$, and $\rho_{MAX} = 0.895$.

Ten random samples of $n = 25, 50, 100,$ and 200 test-takers were sampled from the original dataset. In each of the 4×10 datasets, 36 tests were produced by varying the scale of the items ($df(g)$ = number of categories in the scale – 1) and the score ($df(X)$ = number of categories in the scale – 1) as well as the number of items and their difficulty levels. The polytomous items were constructed from the original binary items by summing relevant sets of items. For example, a test with 20 binary items was divided into 4 subsets: ten items with the scale 0–2, five items with the scale 0–4, four items with the scale 0–5, and two items with the scale 0–10. Consequently, the number of binary items is remarkably higher than other items, the items with wide scale are sparse (see Table 3), and the test items and test scores are partly related to each other.

The dataset of 14,880 items from 1,440 partly related tests consisted varying number of test-takers ($n = 25, 50, 100,$ and 200), items ($k = 2–30$), categories in the items ($df(g) = 1–14$) and in the score ($df(X) = 10–27$), the average difficulty levels ($\overline{p} = 0.50–0.76$) and the lower bound of reliabilities ($\rho_a = 0.55–0.93$). Aside of the estimates, a simple and straightforward statistic is calculated to assess the behavior of the proposed $R_{AC}$ and ACERs: 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. In tables and figures, these variables are named as $dR_{it}$, $dR_{AC}$, and $d\alpha_{RAC}$. The dataset of individual items including several indicators of item–score association is available at http://dx.doi.org/10.13140/RG.2.2.17594.72641. The dataset of reliabilities is available at http://dx.doi.org/10.13140/RG.2.2.27971.94241. Selected basic statistics of the dataset of the test items are collected in Table 3.
Because the true population correlation is unknown in the real-life datasets, all estimators compete in their own race: sample $R_{it}$ is compared with population $R_{it}$ and sample $R_{AC}$ with population $R_{AC}$. The traditional estimators, alpha, theta, omega and rho with their traditional weight factor and score variables, are used as references, and a MCERs based on $R_{AC}$ are used as a benchmark for the ACERs. In ACERs and MECRs, the manifestation of latent trait is the raw score.

First note to make is that, with very small sample size ($n = 25$), both $R_{it}$ and $R_{AC}$ tend to underestimate the population correlation in an obvious manner although $R_{AC}$ less than $R_{it}$ (Figure 1). Second, except the smallest sample size in the simulation, the sample $R_{AC}$ tends to be overestimate the population $R_{AC}$ mildly with small sample sizes and when the scale in the item is wide. This is understood by the fact that, with small sample sizes, the probability to obtain near-deterministic patterns leading to high value of $R_{AC}$ is higher than in the larger population. With items with a narrow scale ($df(g) < 4$) and with sample sizes around $n = 100$ or higher, the possible overestimation is nominal (see also Table 3).

![Figure 1. Average difference between sample and population estimates by Rit and R_{AC} by the sample size and number of categories in the item](image-url)
Third, in the simulation dataset, the average estimates of reliability by DCERs are 0.04–0.07 units of reliability higher than those by the traditional estimators using their traditional linking factor and score variable (Table 4). This follows strictly from the fact that magnitude of the estimates by $R_{AC}$ and $R_{PC}$ tend to be higher than that of those by $R_{it}$.

**Table 4. Average estimates of reliability by selected DCERs**

|        | traditional estimators | ACERs | MCERs |
|--------|------------------------|-------|-------|
|        | $R_{it}$ $\lambda_{(PC)}$ $\lambda_{(ML)}$ $\lambda_{(ML)}$ | $R_{AC}$ | $R_{PC}$ |
| N      | Alpha Theta Omega Rho | Alpha Theta Omega Rho | Alpha Theta Omega Rho |
| 25     | 0.815 0.835 0.819 0.871 | 0.873 0.885 0.914 0.934 | 0.867 0.878 0.911 0.931 |
| 50     | 0.859 0.866 0.862 0.881 | 0.899 0.903 0.928 0.938 | 0.898 0.902 0.929 0.937 |
| 100    | 0.863 0.867 0.865 0.877 | 0.899 0.902 0.928 0.935 | 0.900 0.902 0.930 0.936 |
| 200    | 0.864 0.866 0.865 0.872 | 0.896 0.897 0.926 0.930 | 0.901 0.903 0.931 0.934 |
| Total  | 0.850 0.858 0.854 0.875 | 0.891 0.897 0.924 0.934 | 0.891 0.896 0.925 0.935 |

Fourth, the estimates using $R_{AC}$ and $R_{PC}$ tend to give estimates with largely the same magnitude (Figure 2) and systematically higher than those by the traditional estimators. This seems to refer to the phenomenon that both DCERs refer to the same latent reliability which is underestimated around 5–8% by the traditional estimators regardless of the difficulty level of the test items. It seems that, with extreme datasets, the estimates $R_{AC}$ are mildly higher than those by $R_{PC}$. This difference is nominal though. Simulation of more extreme datasets would shed more light in this matter. That the difference in the magnitude of the estimates in the simulation by the traditional estimators and DCERs is not as dramatic as in Cases 1 and 2 is caused by the fact that the tests in simulation do not allow to prepare tests of extreme difficulty. In the simulation, obtaining tests with extreme difficulty level would have required very short tests using only items with extreme difficulty levels.

Fifth, when it comes to the difference between the sample and population estimates, except the smallest sample size, the ACERs give stable estimates; the differences between the sample and population estimates are subtle. As an example, when using alpha as the base and $R_{AC}$ as the linking factor, the deviance between the sample and population estimates is, on average, 0.001–0.003 units of reliability depending on sample size (Table 5). If the sample size is $n = 200$ (or higher), difference is less than 0.001 units of reliability.
Figure 2. Average estimates of reliability by selected DCERs by the test difficulty

Table 5. Average difference between sample and population by dAlphaRAC

| difference between sample and population | Std. Deviation of the difference | N       |
|----------------------------------------|---------------------------------|---------|
| df(g)                                  | n = 25  | n = 50  | n = 100 | n = 200 | n = 25  | n = 50  | n = 100 | n = 200 | n = 25  | n = 50  | n = 100 | n = 200 |
| 1                                      | -0.026  | -0.002  | 0.001   | 0.000   | 0.034   | 0.016   | 0.011   | 0.008   | 80      | 80      | 80      | 80      |
| 2                                      | -0.035  | -0.002  | 0.002   | 0.000   | 0.046   | 0.022   | 0.015   | 0.011   | 62      | 60      | 60      | 60      |
| 3                                      | -0.035  | -0.001  | 0.002   | 0.001   | 0.058   | 0.024   | 0.015   | 0.012   | 49      | 40      | 40      | 40      |
| 4                                      | -0.038  | -0.002  | 0.006   | 0.000   | 0.060   | 0.028   | 0.023   | 0.012   | 33      | 31      | 30      | 30      |
| 5                                      | -0.034  | 0.009   | 0.007   | 0.005   | 0.066   | 0.031   | 0.023   | 0.013   | 32      | 31      | 22      | 20      |
| 6                                      | -0.018  | 0.016   | 0.000   | 0.001   | 0.052   | 0.033   | 0.015   | 0.011   | 28      | 24      | 22      | 20      |
| 7–14                                   | 0.008   | 0.010   | 0.006   | -0.001  | 0.035   | 0.037   | 0.024   | 0.017   | 76      | 94      | 106     | 110     |
| Total                                  | -0.023  | 0.003   | 0.003   | 0.000   | 0.051   | 0.029   | 0.019   | 0.013   | 360     | 360     | 360     | 360     |
Figure 3. Average difference between the sample and population by selected DCERs by the sample size and item scale.
Sixth, because of mild overestimation of population $R_{AC}$, the ACERs seem to give mild overestimation of the population reliability, specifically, with small sample sizes, polytomous items, and if using rho as the base (Figure 3). The last is expected because rho alone is known to give overestimations with finite sample sizes (Aquirre-Urreta, Rönkkö, & McIntosh, 2019; see also Figure 3); $R_{AC}$ and rho combined seems to lead to greater overestimation than the other combinations. From this viewpoint, the MCERs based on alpha, theta or omega and using $R_{PC}$ as the linking factor tend to lead to more conservative estimates. However, from the factual estimate viewpoint, $R_{AC}$ and $R_{PC}$ seem to lead largely to estimates of same magnitude (see Figure 2 above).

Seventh, regardless of the width of the scale in the items, the estimates by ACERs seem to bring us nearer the population value than the other estimators except if rho would be selected as the base (Figure 4). When the scale is very wide (more than 7 categories), ACERs tend to overestimate reliability mildly, but the values are still nearer the population value in comparison with traditional estimators and MCERs in comparison. Using rho as the base for ACER is not recommendable because of tendency to produce overestimates with finite samples.

Figure 4. Average estimates of reliability by selected DCERs by the number of categories in the items
Appendix 3. Numerical example of the effect of ACERs to standard error of the measurement (S.E.m)

Within the text, it was referred to a dataset by Metsämuuronen and Ukkola (2019) where a very easy screening test of language skills related to the language used in the test was administered. Only a minority of students with the instruction language as the second language were expected to fail to give correct answers in some of the items. The advance of ACERs in estimating the standard error of the measurement (S.E.m) may be notable in these kinds of datasets where the item difficulties are extreme leading to an ultimately non-normal score (see Case 1 above).

Assume we are interested in S.E.m of the raw scores in the test. Maximum number of points are 11, mean 10.57, and standard deviation 0.875. Item-wise statistics are collected in Table 6.

| Item (g) | Max | Mean | p | VAR(g) | Rit (obs.) | Rit (max.) | RACX | Rit × SD | RACX × SD |
|---------|-----|------|---|--------|------------|------------|-------|--------|----------|
| g1      | 1   | 0.964 | 0.964 | 0.035 | 0.350 | 0.635 | 0.551 | 0.065 | 0.103 |
| g2      | 1   | 0.984 | 0.984 | 0.016 | 0.268 | 0.548 | 0.490 | 0.034 | 0.062 |
| g3      | 1   | 0.992 | 0.992 | 0.008 | 0.288 | 0.478 | 0.603 | 0.025 | 0.053 |
| g4      | 1   | 0.909 | 0.909 | 0.082 | 0.455 | 0.754 | 0.603 | 0.131 | 0.173 |
| g5      | 2   | 1.785 | 0.892 | 0.372 | 0.747 | 0.811 | 0.921 | 0.455 | 0.562 |
| g6      | 1   | 0.985 | 0.985 | 0.015 | 0.258 | 0.541 | 0.477 | 0.032 | 0.058 |
| g7      | 2   | 1.973 | 0.986 | 0.045 | 0.329 | 0.579 | 0.568 | 0.069 | 0.120 |
| g8      | 2   | 1.981 | 0.990 | 0.028 | 0.376 | 0.553 | 0.680 | 0.063 | 0.115 |
| SUM     |     | 0.600 |     |       | | | 0.875 |    | 1.245 |

Using Eq. (1), we get the traditional estimate of reliability as \( \alpha = \frac{8}{8-1} \left( 1 - \frac{0.600}{0.875^2} \right) = 0.246 \).

Similarly, using Eq. (11), we get the attenuation-corrected estimate of reliability as \( \alpha_{RAC} = \frac{8}{8-1} \left( 1 - \frac{0.600}{1.245^2} \right) = 0.700 \). Applying the standard formula of the standard error of the measurement \( (S.E.m = \sigma_s = \sigma_x \sqrt{1-REL}) \), the traditional alpha suggests that S.E.m is 0.875 × \sqrt{1-0.25} = 0.760 points while the deflation-corrected alpha suggests only 0.875 × \sqrt{1-0.700} = 0.479 points. Hence, the measurement error increases (at least) by 58% = (0.76 – 0.48)/0.48) because of technical reasons if we use the traditional alpha instead of deflation-corrected alpha. Alternatively, because of technical reasons, a magnitude of a rough general estimate of the measurement error may decrease by 36% or more if we use a deflation-corrected alpha instead of the traditional alpha. The “at least” comes from the fact that some other deflation-corrected estimators of reliability such as alpha combined by \( R_{PC} \), \( G \) and \( D \) give estimates \( \alpha_{RPC} = 0.790, \ \alpha_G = 0.850, \ \) and \( \alpha_D = 0.856 \) respectively for this specific dataset (see Metsämuuronen, 2022a). Then, it seems that, in comparison with other types of DCERs, ACERs combined with \( R_{AC} \) give conservative estimate in the case of tests with extreme item difficulty. In any case, selecting wisely estimators of reliability that produce estimates being nearer the true reliability value may give us a notable advance in assessing the accuracy of the test scores.