Behaviormetrika
https://doi.org/10.1007/s41237-025-00279-0

ORIGINAL PAPER

Dimension-corrected discrimination index

Jari Metsämuuronen[1]

Received: 27 December 2022 / Accepted: 20 September 2025

Abstract

The Discrimination Index (*DI*) or upper-lower index is a simple and robust, classical non-parametric shortcut to evaluate the item discrimination power or item validation in the practical testing settings. Empirical results show that the *DI* has two shortcomings: first, the traditional *DI* with the 27% cut-off appears to underestimate for the item discrimination power and, second, for polytomous items, the *DI* behaves illogically compared to the better performing benchmarking estimators. Therefore, two suggestions are made to modify the procedures when using *DI*. First, it would be preferable to use the 20% cut-off of instead of 27%; this seems to tend to give estimates that are closer to the better performing association estimators than the point-biserial correlation which is known to underestimate item-score association. Second, two simple modifications of *DI*, dimension-corrected *DI* (DI_2), are proposed: $DI_2Log = DI_{20\%} + 0.146LN\,(R-1)$ and $DI_2Lin = DI_{20\%} + 0.05\,(R-2)$, where $DI_{20\%}$ refers to the observed value of *DI* with a 20% cut-off, *LN* refers to the natural logarithm, and *R* is the number of categories in the item. Based on the training dataset and external datasets, DI_2 seems safe to use at the difficulty levels between p=0.20–0.80 and when the number of categories in the item does not exceed 7–8.

Keywords Kelley's discrimination index · Upper-lower index · Generalized discrimination index · Item analysis · Item–total correlation · Dimension correction

Communicated by Kentaro Kato

✉ Jari Metsämuuronen
jari.metsamuuronen@gmail.com

[1] Faculty of Mathematics and Natural Sciences, Turku Research Institute for Learning Analytics (TRILA), University Of Turku, FI-20014, Turku, Finland

Published online: 01 November 2025

1 Introduction: discrimination index in historical perspective

1.1 Discrimination index and generalized discrimination index

The discrimination index (*DI*; Long and Sandiford 1935; Kelley 1939), also known as Kelley's DI (e.g., Metsämuuronen 2017, 2020a) or upper-lower index (ULI; e.g., Çelen and Aybek 2022; Martinková et al. 2017), is one of the classical indices reflecting the discrimination power of a test item and can therefore be used as a rough tool for test item validation. In this context, the concept of 'item discrimination power' (IDP) is often loosely defined as the efficiency of a test item to discriminate between lower and higher performing test takers (see discussion in, e.g., Lord et al. 1968; ETS 2022; Liu 2008; Macdonald & Paunonen 2002; Metsämuuronen 2020b; Moses 2017). From this perspective, *DI* differs from many other indices: while most indices used for the same purpose estimate something specific—for example, coefficients based on covariation estimate the true correlation in the population, and coefficients based on probability estimate the true probability in the population—*DI* does not seem to *estimate* commonly known entity. However, the formula itself refers to classical probability (observed/maximum). Nevertheless, this probability is underdefined in literature (cl. *D*, *G*, and *Tau* which estimate the probability that two random pairs are in the same order in two variables).[1] Therefore, technically speaking, it may be a misnomer to say that *DI* is an estimator of item discrimination or item-score association. However, because it roughly indicates how effective an item is at separating test takers with the lower level of the trait from those with the higher level, we can loosely say that *DI* 'estimates' item discrimination power or item-score association. Another statistical peculiarity of the *DI* is that it does not (yet) have a sampling distribution, so we cannot make statistical inferences about its values.

Unlike other indices of item discrimination power, *DI* uses a special procedure in which only the extreme cases of the ordered data set are used in the analysis. Discussion of different cut-offs of the extreme cases based on the ordered score was active in the 1930 s and 1940 s (e.g., Forlano and Pinter 1941; Kelley 1939; Long and Sandiford 1935; see also Johnson 1951) and later in the 1960 s and 1970 s (e.g., Cureton 1966a, b; D'Agostino, and Cureton, 1975; Ebel 1967; Feldt 1963; Ross and Lumsden 1964; Ross and Weitzman 1964). The cut-offs usually suggested in the literature are either 25% of the extreme cases (e.g., D'Agostino & Cureton 1975, Mehrens, and Lehmann, 1991; Metsämuuronen 2017) or 27% of the extreme cases (e.g., Ebel 1967; Feldt 1963; Johnson 1951; Kelley 1939; Ross and Weitzman 1964). Of these, the 27% cut-off is more mathematically justified because it maximises the difference in the population with normally distributed scores when the average difficulty of the score is 50% (see Kelley 1939; Wiersma and Jurs 1990). However, Kelley (1939) specifically pointed out that if the score is not normally distributed with $p(X) = 0.50$, then the cut-off that maximises the population value is something other than 27%, although he did not specify the cut-offs. In real life situations with small

[1] An anonymous reviewer suggested that DI may estimate the population quantity, since the upper part of the formula fulfils the definition of population quantity.

sample sizes, it may not be possible to find a 27% cut-off. Then any cut-off close to 25–27% can be used.

Other options have also been explored. For example, Forlano and Pinter (1941), after examining cut-offs of 50%, 33%, 27%, 16% and 7%, concluded that although none was absolutely superior to the others, 27% provided an effective estimate. Liu (2008) compared the cut-offs of 50%, 33%, 27% and 10% and concluded that the 33% cut-off would be the most recommended for use in classroom item analysis. Beuchert and Mendoza (1979) suggest that if the dataset is large enough, 50% and 30% cut-offs can give broadly similar results to 27%. Based on the results of this article, it appears that both Liu and Beuchert and Mendoza use the item-total correlation (*Rit*) as a benchmarking measure of *DI*. However, Metsämuuronen (2022c; see also, e.g., McGrath and Meyer 2006; Ruscio 2008) has pointed out that *Rit* remarkably underestimates the true association between an item and the score.

In the last 20 years, *DI* has not been actively discussed in the methodological literature. However, Liu (2008) used *DI* as one of the coefficients in a general comparison of several indices of discrimination, Bazaldua et al. (2017), Kelley et al. (2002Metsämuuronen 2022b); Tristan (1998) connected *DI* with item response theory (IRT) modelling, Batanero (2007) linked *DI* to Bayesian methodology, Celen and Aybek (2022) used *DI* as a benchmark coefficient for their new item discrimination power estimators, and Metsämuuronen (2020a, 2022a) discussed the generalised discrimination index. Even though the theoretical discussion has been relatively settled in recent years, *DI* may be widely used in practical testing settings in educational and medical fields (see a collection of recent literature in Metsämuuronen 2020a). This is expected because *DI* has been actively promoted as a suitable and simple tool to quickly get a rough idea of how well the binary test items would discriminate between the lower and higher achieving test takers (see, chronologically, Johnson 1951; Ebel 1954a, b, 1972, 1979; ETS 1960; Nitko et al. 1984; Ebel and Frisbie 1986; Mehrens and Lehmann 1973, 1991; Wiersma and Jurs 1985, 1990; Metsämuuronen 2017). The use of *DI* as a coarse tool in item analysis is justified because the index is easy to calculate in practical measurement settings without complicated software, it is more stable for binary items than another traditional estimator, the item-total correlation (*Rit*; cf. Metsämuuronen 2020a), and unlike *Rit*, it can (correctly) indicate the deterministically discriminant dataset and then $DI=1$. In addition, the formula for *DI* can be used to analyse the effectiveness of each distractor in a multiple-choice question (Nitko et al. 1984).

Since *DI* is traditionally restricted to binary settings and fixed cut-offs, Metsämuuronen (2020a) introduced the generalised *DI* (*GDI*), which can be used with binary and polytomous items and uses all possible cut-offs. The formulas will be discussed later. Metsämuuronen's procedure of exhaustive splitting (PES) has led to a new type of graphical tool, the cut-off curve, for visual diagnosis of items (Metsämuuronen 2022a), and the procedure has been used to estimate the bias-corrected difficulty level (Metsämuuronen 2022b). The characteristics of the *GDI* and thus the *DI* for the polytomous items have not been widely studied (see, however, (Çelen, and Aybek, 2022). This study aims to shed light on this issue. In the following empirical section, the abbreviation *GDI* is used when both the binary and polytomous items are of interest and when several cut-offs are compared with some benchmark estimators, and *DI*

is used when a specific cut-off is of interest regardless of the binary or polytomous nature of the scale in the item.

1.2 Computational forms of DI and GDI

Traditionally, *DI* is calculated using the following procedure. Consider a test with *N* examinees ranked by score (*X*). The candidates are divided into an upper part (*U*) consisting of the highest scoring candidates and a lower part (*L*) consisting of the lowest scoring candidates. Using this notation and assuming a binary item, *DI* can be expressed as follows

$$DI = \frac{R^U - R^L}{\frac{1}{2}T} = 2\left(p^U - p^L\right) = 2\left(p - 2p^L\right) \tag{1}$$

(e.g., Mehrens and Lehmann 1991; Metsämuuronen 2017, 2020a), where R_U and R_L refer to the number of correct answers in the upper and lower parts of the ordered dataset, and *T* refers to the total number of test takers in the two parts together. Consequently, p_U and p_L refer to the proportions of correct answers in the top and bottom parts of the dataset, and *p* is the proportion of correct answers in these sections combined. Note that above the *p*s are calculated using the combined number of cases in both parts (*T*). Alternatively, the number of cases in one of the partitions (½*T*) could be used, and this leads us to the form

$$DI = p^U - p^L \tag{2}$$

(e.g., Johnson 1951; Liu 2008; Martinková and Drabinová 2018; Martinková et al. 2017; see also Çelen and Aybek 2022), where p^U and p^L are the proportions of the correct answers in the upper (*U*) and lower (*L*) groups *independently*. Equation (2) leads us to the possibility of calculating *DI* also for the polytomous items (see Çelen and Aybek 2022; Martinková et al. 2017), although *DI* is usually calculated for binary items. The polytomous cases can also be included by using the form of *GDI*:

$$GDI_a = \frac{R_a^U - R_a^L}{\frac{1}{2}T_a} = 2\left(p_a - 2p_a^L\right) \tag{3}$$

(Metsämuuronen 2020a), where the subscript *a* refers to the symmetric cut-off used in the estimation either in percentage (e.g., *a*=25% or *a*=27%), the number of cases in the cut-off *a* (*a*=5) or rank-order of the cut-off (*a*=5=5th case in the ordered dataset). In Eq. (3), instead of the *number* of correct answers, R_a^U is the *sum of the observed item scores* of the test takers from the highest to the a^{th} highest case, and, in parallel, R_a^L is the sum of the observed item scores of the test takers from the lowest to the a^{th} lowest case. Similarly, *T* refers to the *maximum possible sum minus the minimum possible sum* of the observed item scores of the test takers in the specific cut-off *a*. In the achievement testing where the minimum value in the item is usually 0, this is not a problem; *T* refers to the maximum possible sum of the respondents in the specific cut-off *a*. However, this re-conceptualization is obvious for types of

scales that do not start from 0, such as a 5-point Likert scale anchored to the values 1–5.

Both Eqs. (2) and (3) embed a seed for very conservative values for polytomous items. As an example, suppose that 20 test takers form a sequence of item responses as follows, ordered from the lowest to the highest score, untraditionally, in horizontal way: 00000000001111111111. If the item was binary, the traditional *DI* with a 25% cut-off (a=5) would give the value of$DI = p^U - p^L = 1 - 0 = 1$, i.e., the *DI* would indicate that the item perfectly discriminates between the lower and higher scoring test-takers. However, if the same pattern is obtained, but the maximum value in the item is 5, $DI = p^U - p^L = (5 \times 1)/25 - (5 \times 0)/25 = 1/5 - 0 = 0.20$, which is traditionally taken as a cut-off value to indicate that the item is in danger of being rejected from a test because of low item discrimination power (see Ebel 1965). The same estimate would be obtained if the pattern were, for example, 44444444445555555555: $DI = (5 \times 5)/25 - (5 \times 4)/25 = 5/5 - 4/5 = 0.20$. *DI* is therefore sensitive to the spacing of the observed item scores. Only for the pattern in which all test takers in the lower part receive the *lowest* possible item score and all cases in the upper part receive the highest possible item score does *DI* reach a value of 1. In contrast, the better-behaving benchmarking estimators discussed later are not bound by the spacing between the observed item scores. Therefore, the better-performing estimators would detect perfect discrimination in both of the cases. For *DI*, this is not an issue in the binary settings, but it becomes an issue in the polytomous settings.

Finally, there are (at least) two types of generalised *DI* other than the one proposed by Metsämuuronen (2020a). Brennan's *B* (Brennan 1972) is generalised in the sense that the cut-off need not be symmetrical. However, *B* is restricted to dichotomous items and uses fixed cut-offs. Harris and Wilcox (1980) showed that Brennan's *B* is algebraically equivalent to Peirce's *theta* discussed by Goodman and Kruskal (1959). Also, Whitney and Sabers (1971) discuss ways of using *DI* in polytomous elements. Neither is widely used with polytomous items. The reason for this may be that the traditional way of calculating *DI* for polytomous items leads to an obvious and serious underestimation of item discriminability, as will be seen in the empirical section.

1.3 Some benchmarking estimators for DI and GDI

There are many indices of item discrimination power. When it comes to comparing the different indices, three waves of research can be distinguished. The first wave of research was mainly interested in comparing the indices within classical test theory *in general*, from the point of view of their usefulness in practical testing situations. Some studies in this regard have been conducted by, for example, Beuchert & Mendoza, 1979, Cureton 1966a, b; Englehart (1965); ETS (1960); Hales (1972); Oosterhof (1976); Wolf (1967); see also more recently, for exampleLiu (2008); Metsämuuronen (2020b).

With the rise of Rasch modelling (Rasch 1960) and item response theory (IRT) modelling in the 1960 s and 1970 s (see, e.g., Lord et al. 1968 and the discussion of its early stages in Rasch, 1972), the second wave of researchers was interested in comparing estimators within classical test theory and IRT modelling. These types of

studies were conducted by, for example, Cook et al. (1988), Fan (1998), Hambleton and Jones (1993), Kelley et al. (2002), Lawson (1991), Macdonald and Paunonen (2002), Ndalichako and Rogers (1997), Shannon and Cliver (1987); see also more recently, for example, Kohli et al. 2015). For example, Kelley et al. (2002) compared the behaviour of *DI* estimates with point-biserial correlation and Rasch item discrimination.

The third wave of research has not been interested in the estimators of *association* per se but, instead, the enhanced estimators of *reliability* because the item-score correlation (*Rit*), one of the indices for the item discrimination, is embedded in the most commonly used reliability formulae such as the coefficients alpha, theta, omega and rho, or maximum reliability (see e.g., Metsämuuronen 2022f, 2022h, 2022i). In this regard, Metsämuuronen (e.g., 2020b,2021a, 2022c) has been active in investigating alternative item-score association estimators to replace *Rit*, which is known to radically underestimate the correlation between the item and the score variable. This has led to a new paradigm related to attenuation or deflation correction in the estimators of reliability including the attenuation-corrected estimators of reliability (ACER; Metsämuuronen 2022f) and deflation-corrected estimators of reliability (DCER; Metsämuuronen 2022f, 2022h; Metsämuuronen 2022i; see also ordinal alpha and theta by Zumbo et al. 2007; see also Gadermann et al. 2012). In ACERs the traditional *Rit* and factor loadings are replaced by attenuation-corrected estimators, and in DCERs, they are corrected by entirely different estimators later referred to as "better-performing" estimators of item-score association. As part of this paradigm, some new estimators have been developed by improving the worse-behaving estimators of association: Dimension-corrected[2] *D* (D_2; originally Metsämuuronen 2020c and corrected in 2021a), dimension-corrected *G* (G_2; Metsämuuronen 2021a), attenuation-corrected product-moment correlation (*RAC*; Metsämuuronen 2022f), and attenuation-corrected *eta* (*EAC*; Metsämuuronen 2022e). These are briefly discussed in what follows (see also their computing in Appendix 1).

Overall, the most recent wave has compared item-score association estimators in terms of their ability to reflect the true association between the latent trait and the item. Particular attention has been paid to the extent of technical or mechanical error in the estimates, which is also discussed under the topic of deflation. In particular, Metsämuuronen (2022c, d) notes radical differences between estimators in reflecting the true association between the item and the latent variable. Some estimators contain a significant amount of deflation. This is common for estimators such as *Rit*, also known as point biserial and point polyserial correlation, and hence item-rest correlation (*Rir*, Henrysson 1963), Spearman rank correlation (*RS*; Spearman 1904), distance correlation (*dRit*); Székely et al. 2007) and coefficient eta (Pearson 1903, 1905), all of which use the formula for the product-moment correlation coefficient (Bravais 1844; Pearson 1896), and the non-parametric estimator Kendall *tau-b* (Kendall 1945). In practical testing settings, this means that the magnitude of the estimates from these

[2] In the term "dimension-corrected", the term "dimension" comes from cross tables with degrees of freedom related to the number of categories, i.e., dimensions of the cross table. Practically all the nonparametric association estimators discussed in this paper (*D*, *G*, D_2, G_2, *Tau*) as well as *RPC* are analysed using the cross tables. There, the word "dimension" refers to the number of cells in each direction.

estimators is radically underestimated relative to the true association between item and score. This is particularly true for items of extreme difficulty. Algebraic reasons for underestimation are discussed, for example, by Metsämuuronen (2016, 2022d for *Rit*; 2017, 2022g for *Rir*; 2022e for *eta*; 2021b for *Tau-b*).

In contrast, certain estimators have been shown to behave much better than those discussed above in terms of deflation in the estimates (see, for example, simulations in Metsämuuronen 2021a, 2022c). These better performing estimators include the polychoric correlation coefficient (*RPC*, Pearson 1900, 1913), the r-bireg and r-polyreg correlation coefficients (*RREG*; cf. Livingstone & Dorans, 2004, Moses 2017) as better alternatives to the traditional biserial and polyserial correlation coefficients, the Somers delta (*D*; Somers 1962) and Goodman-Kruskal gamma (*G*; Goodman and Kruskal 1954) with binary items and their dimension-corrected versions D_2 and G_2 for binary and polytomous settings, as well as attenuation-corrected *Rit* (*RAC*; Metsämuuronen 2022f) and attenuation-corrected eta (*EAC*; Metsämuuronen 2022e). It seems that *DI* belongs partly to the group of worse estimators and partly to the group of better estimators. A numerical example of their behaviour illustrates the differences between the worse and better estimators.

1.4 Numerical example of the traditional DI compared with selected item-score association estimators

Consider a hypothetical dataset as in Table 1. With $n=15$ cases, the closest cut-off to the 27% is $a=4$ cases, resulting in a cut-off of 26.6%. These cases are highlighted in Table 1. The dataset contains two items of interest, a binary item ($g1$) and a 0–4 scaled polytomous item ($g2$), and a score (X) without tied cases (comprised of several other items). Both items are deterministic, i.e., after ordering the respondents by score, the item can deterministically discriminate the lower scoring respondents from the higher scoring respondents. In these kinds of settings, $RPC \approx RREG \approx G = G_2 = D = D_2 = RAC = EAC = 1$, i.e., these estimators can detect the deterministic pattern through the perfect correlation either asymptotically ($RPC \approx RREG$) or exactly. In contrast, *Rit* (and the related *RS*, *eta*, and *dRit*) cannot detect the deterministic pattern. In particular, *DI can* detect the deterministic pattern in the binary settings when the item difficulty is $0.27 > p > 0.73$ when the 27% cut-off is chosen, and $0.25 > p > 0.75$ when the 25% cut-off is chosen. This characteristic places it in the group of better performing estimators. However, for items of very extreme difficulty, *DI* can give an obvious and radical underestimation of item discrimination.

For Table 1, the estimates by *dRit* are calculated by using the R package *Energy* by Rizzo and Székely (2022).[3] The estimates of *Rit*, *eta*, *RS*, *G*, *D*, and *RREG* are calculated using the IBM SPSS software package and *RPC*, *DI*, D_2, G_2, *RAC*, and *EAC* are calculated using the MS Excel software package either strictly (*DI*), by extending an existing procedure (*RPC*)[4], or by knowing the estimates of *G*, *D*, *Rit*, and *eta*. The

[3] Sincere thanks to Counselor or Evaluation, Jukka Marjanen from Finnish Education Evaluation Centre (FINEEC), for his support in the calculation of *dRit*.

[4] Zaiontz's algorithm (2022) is slightly improved because the pattern in the data set is deterministic. This will be discussed later. With respect to *RPC*, different software packages solve the challenge of determin-

Table 1 Hypothetic dataset for a comparison of DI and benchmarking estimators of association

ID	g_1	g_2	X
1	**0**	**0**	**3**
2	**0**	**1**	**4**
3	**0**	**1**	**8**
4	**1**	**2**	**9**
5	1	2	10
6	1	2	15
7	1	2	16
8	1	3	18
9	1	3	20
10	1	3	24
11	1	3	28
12	**1**	**4**	**29**
13	**1**	**4**	**32**
14	**1**	**4**	**34**
15	**1**	**4**	**35**
$DI_{27\%}$	0.750	0.750	
Rit	0.664	0.946	
eta	0.664	0.962	
RS	0.694	0.972	
dRit	0.675	0.944	
RREG	0.969	1.000	
RPC	1.000	1.000	
G	1	1	
D	1	1	
G_2	1	1	
D_2	1	1	
RAC	1	1	
EAC	1	1	

formulation and syntax will be discussed in Appendix 1. As an example, in the case of item g_2, the estimate by $DI_{27\%}$ with $a=4$ is calculated as $16/16 - 4/16 = 0.750$.

In particular, the magnitudes of the estimates by *DI* are quite high (0.750); the identical estimates for item g1 and g2 are a coincidence. If we take the estimates by *RPC*, *RREG*, *G*, *D*, D_2, G_2, *RAC*, and *EAC* as close reflections of the true association—after all, they correctly or closely indicate the deterministic pattern related to the true association—the magnitude of the estimates by *DI* is on the one hand reasonably close to the true association, but on the other hand significantly away from the true association; the deflation is 25% in both cases. The behaviour of *DI* and *GDI* in polytomous settings is largely unknown. It is clear that *DI* does not always detect the deterministic pattern in polytomous items, but we do not know how it behaves in polytomous settings in general. This study sheds some light on this.

istic patterns in different ways. For example, the R package *polycor* (Fox and Dusa 2022), gives estimates of 0.970 and 0.947 while the manual calculation gives a closer approximation to perfect correlation. The reason for the difference may also lie in the different algorithms used in the different packages. Technically, *RPC* cannot achieve the perfect correlation *RPC*=1 unlike *D* or *G*, for example.

2 Research questions

In what follows, an empirical study is carried out in two phases. In the first phase, the characteristics of *DI* and *GDI* are compared with benchmark estimators, specifically in the polytomous empirical settings. In this respect, one benchmark for apparent underestimation is *Rit*, which is known to underestimate the true association whenever the scales of two variables are not identical (see algebraic discussions in, for example, Metsämuuronen 2016, 2022d and simulations in, for example, Martin 1973, 1978; McGrath and Meyer 2006; Metsämuuronen 2021a, 2022c, Olsson 1980; Ruscio 2008). Other benchmarks are various types of item-score association estimators that have performed significantly better in simulations: *RPC*, *RREG*, G_2, D_2 and *RAC*. The motivation for selecting these latter estimators will be discussed later.

In the first stage, it will be seen that the *DI* estimates have two shortcomings or weaknesses. First, the traditional *DI* with the 27% cut-off tends to underestimate the *IDP*, especially for polytomous items, but also to some extent for binary items, compared to the better performing estimators. Secondly, the *DI* behaves very differently, if not illogically, with polytomous items compared to the other estimators. Therefore, in the second phase, a simple modification of *DI* is proposed, dimension-corrected *DI* (DI_2). This modification brings the estimates of *DI* closer to the estimates of the better performing estimators. Some relevant characteristics of DI_2 are studied using a simulation with a real dataset on the one hand, and a theoretical dataset that detects the possible technical deflation in the estimates on the other hand.

3 Methods

3.1 Statistical model related to DI and GDI in comparison with the other estimators

Let us assume that the observed item scores in item g (y_i) with $r=1, \ldots, R$ binary or ordinal categories and a metric score variable X (x_j) with $c=1, \ldots, C$ ordinal, interval, or pseudo-continuous categories, share the common latent variable θ (e.g., "proficiency in Y" or "attitude towards Z") and $R<<C$. Let us denote the threshold values of θ for each category in g by γ_i and in X by τ_j so that $g=y_i$, if $\gamma_{i-1} \leq \theta < \gamma_i$, $i=1, \ldots, R$ and $X=x_j$, if $\tau_{j-1} \leq \theta < \tau_j$, $j=1, 2, \ldots, C$, and $\gamma_0 = \tau_0 = -\infty$ and $\gamma_R = \tau_C = +\infty$. With the traditional item-score association estimators based on covariation (e.g., *Rit*, *RPC*, *RREG*, *RAC*, *EAC*) or probability (e.g., *Tau-b*, *G D*, G_2, D_2), this model can be illustrated as in Fig. 1. In Fig. 1, n_{ij} refers to the number of cases in each cell, and the form mimics the inversed cross-table; a positive association is seen as a negative trend.

The statistical model associated with *DI* differs radically from that associated with the estimators above in that, first, the intermediate categories of X are not used at all, and, second, the categories in X within the range of the lower extreme (L) and upper extreme (U) and L are truncated into one "category", and, third, *DI* consequently does not use the score in any other way than simply to form the order of the test takers. Technically, g and X are related to θ such that $g=y_i$, if $\gamma_{i-1} \leq \theta < \gamma_i$, $i=1, \ldots, R$ and

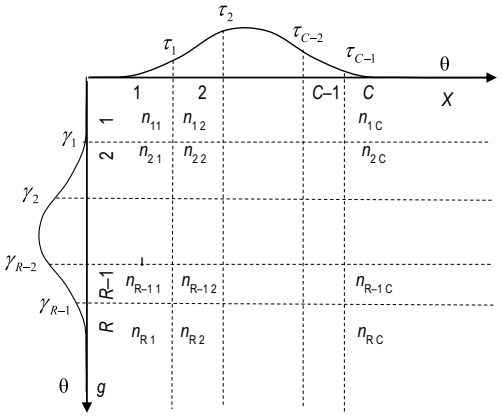

Fig. 1 Statistical model related to a common latent variable θ manifested in two different observed variables g and X

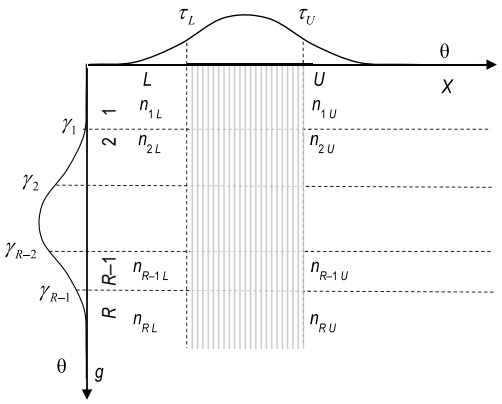

Fig. 2 Statistical model for DI and GDI

$X=L$, if $\theta < \tau_L$ and $X=U$, if $\theta > \tau_U$. This is illustrated in Fig. 2. In particular, in the traditional settings with *DI*, $R=2$ (with $y_i = 0$ and 1), and the cut-offs can be fixed (in *DI*) or we can use several or all cut-offs in the analysis (*GDI*).

3.2 Dataset used in the empirical section

A simulation dataset based on a real-world dataset is used to examine the characteristics of *DI* and *GDI* in comparison with the benchmark association estimators on the one hand and in the polytomous settings on the other. The simulation dataset is based on 4,023 nationally representative test takers of a mathematics test with 30 binary test items. A random sample of 1,440 compilations of items with different sample sizes and difficulty levels was drawn from the original dataset, resulting in 14,880 estimates of several item-score association estimators. Estimates by *DI* were computed with 27% and 20% cut-offs for all items and for 4%, 8%, 12%, 16%, 24% and 28% for 8,160 items; the 6,720 estimates omitted were from binary items. The dataset is available at https://doi.org/10.13140/RG.2.2.24874. in SPSS format and at https://doi.org/10.13140/RG.2.2.14808.16648 in CSV format.

This training dataset was formed from the original binary dataset of $n=4{,}023$ test takers (FINEEC 2018). In this original dataset, item discrimination ranged $0.332 < Rit < 0.627$ with mean $\overline{Rit}=0.481$, the difficulty levels of the items ranged $0.24<p<0.95$ with mean $\bar{p}=0.63$, and the lower bound of the reliability was $\alpha=0.885$. Ten random samples of $n=25$, 50, 100, and 200 test takers were drawn from this dataset. These represent different sample sizes typical of real-life testing settings with educational settings relevant to *DI*: $n=25$ may be a typical sample size for classroom testing and $n=200$ may be the sample size for a lecture to a large group of students or a common test in a school for all students of the same age. In each of the 10×4 data sets, 36 shorter tests were constructed with 20, 21, 22, 24, 26, 27, 28, and 30 binary items. In each of the $36\times8=288$ independent sets, the polytomous items were constructed as sums of the original binary items. For example, a test of 28 points was divided into three sets: one test of 28 binary items ($k=28$, $R=2$, $C=28$), nine tests by summing every ninth binary item ($k=9$, $R=3$, $C=28$), and three tests as summing every third item ($k=3$, $R=9$, $C=28$). Similarly, the sets of 21, 22, 26, and 27 items produced three such "parallel" sets of tests, while sets of 20, 24, and 30 items produced 5, 7, and 7 test sets, respectively.

As a result, $36\times40=1{,}440$ partially related tests with 14,880 test items form a data set with a varying numbers of test takers ($n=25$, 50, 100, and 200) and items ($k=2$–30, $\bar{k}=10.33$, std. dev. 8.621), the average difficulty levels of the items in the test ($p_g=0.08$–0.96, $\bar{p}_g=0.66$. std. dev. 0.107) and of the score ($p_X=0.50$–0.76, $\bar{p}_X=0.66$. std. dev. 0.051), and the number of categories in the items ($df(g)=R-1=1$–14, $\overline{df(g)}=4.57$, std. dev. 3.480) and in the score variables ($df(X)=C-1=10$–27, $\overline{df(X)}=18.06$, std. dev. 3.908). Due to the routine of creating the polytomous items by summing the binary items, it is noteworthy that the items with more than 7 categories ($R=7$–14) are always related to a test with only three or two items with a wide scale, and the items with more than 10 categories are always related to a test with only two items. The lower limits of the reliabilities estimated by coefficient alpha vary from low to high ($\rho_\alpha=0.55$–0.93, $\bar{\rho}_\alpha=0.850$, std. dev. 0.049). This dataset will be referred to as the training dataset. In fact, the modelling was done in two ways. In the alternative reported here, the entire dataset was used to train the model. The result was checked using the traditional verification procedure where two random splits of the dataset ($n=7{,}440$ in both datasets) were formed of which one part was used for training and the other for the verification purposes. In this latter procedure, the result was the same as in the first alternative: the model formed by the training split fits the verification split almost perfectly. This may be due to the peculiarities embedded in the simulation data set.

3.3 Selected benchmarking indices of item discrimination power

Although we have more than 20 traditional item discrimination indices that have been widely discussed in the literature (see, for example, Oosterhof 1976; who compared 19 of them), and several new ones that have been developed after the first wave of studies (such as *dRit*, D_2, G_2, *RAC*, *EAC*, and several alternatives for *Rir*; see, for example, Metsämuuronen 2022 g), only a few of them are widely used, as Liu (2008)

correctly points out. Metsämuuronen (2020b) notes that two of them seem to dominate the options given to practical users: *Rit* and *Rir* are the defaults for the item discrimination indices in the widely used general software packages such as IBM SPSS, STATA and SAS (see IBM 2017; Stata corp., 2018; Yi-Hsin and Li 2015).

It is known that some of the item-score association estimators give underestimates for mechanical or technical reasons (see above), and *Rit* can be used as a benchmark for this. Since *Rit* always underestimates the item-score association when the scales differ, if the magnitude of an estimate from any other estimator is smaller than that of *Rit*, it can be taken as an obvious underestimate.

Some of the better performing options are based on the covariation with the latent trigonometric nature such as *RPC* and *RREG* (for the nature of the estimators and their effect in the estimates, see Metsämuuronen 2022c). Some of the better performing nonparametric estimators are based on the probability with a latent linear nature such as *G* and *D*. These are known to give obvious underestimates with polytomous items (see Metsämuuronen 2020b, 2021a). Therefore, they are not used in in the following, but their dimension-corrected versions G_2 and D_2 with a semi-trigonometric nature are used instead. In the binary setting, $G_2 = G$ and $D_2 = D$. It is known that D_2 usually gives more conservative estimates than G_2; this follows from the relationship of *D* and *G* (see e.g., Metsämuuronen 2021b). Another pair of new estimators of either trigonometric or linear nature are the attenuation-corrected *Rit* (*RAC*) and the attenuation-corrected *eta* (*EAC*). Of these, *RAC* is used as the benchmark. The computational forms and syntax of these estimators are discussed in Appendix 1.

4 Results

4.1 DI behaves radically differently in comparison with the benchmarking estimators

The first takeaway from the analysis is that *DI* behaves strikingly differently from the benchmark estimates of item-score association. Whereas the other indices approach the value of 1 as the number of categories in the item increases, the values of *DI* tend to be "constant" as the number of categories increases (Table 2; Fig. 3). The reason for the difference between *DI* and the association estimators may be that *DI* does not use the actual score values in the calculation process (see Fig. 2 and related discussion).

That the correlation approaches (or should approach) the value of 1 is a unique and obvious feature of the item-score association when we think of a 'test' with only one item: the correlation between the item and the 'score' formed by that item would, of course, be perfectly 1. The item-score association approaches the perfect 1 the fewer items there are in the test and the more categories there are in the items; this does not make sense outside of measurement modelling. The theoretical reason for this unique phenomenon is that the latent variable θ is common to both items and score (see Fig. 1), and the more similar items and score are, the closer the correlation approximates the perfect one. Later, a modification of *DI* is proposed to bring its tendency closer to the better performing estimators.

Table 2 Average estimates of item–score association by selected estimators

df(g)	Number of estimates	Rit	RPC	RREG	G2	D2	RAC	Mean of RPC to RAC	DI27%	DI20%
1	7,948	0.473	0.6162	0.5892	0.6144	0.5923	0.5863	0.5997	0.5328	0.5981
2	3,056	0.6122	0.6918	0.6697	0.6852	0.6610	0.6746	0.6765	0.5306	0.5886
3	1,390	0.7007	0.7521	0.7366	0.7610	0.7368	0.7457	0.7464	0.5303	0.5881
4	729	0.7659	0.8018	0.7891	0.8202	0.7955	0.8009	0.8015	0.5346	0.5935
5	474	0.8092	0.8347	0.8274	0.8555	0.8327	0.8376	0.8376	0.5353	0.5947
6	366	0.8477	0.8658	0.8589	0.8889	0.8678	0.8706	0.8704	0.5312	0.5880
7	255	0.8776	0.8909	0.8874	0.9132	0.8924	0.8974	0.8963	0.5374	0.5976
8	140	0.9007	0.9096	0.9088	0.9317	0.9131	0.9181	0.9163	0.5323	0.5883
9	160	0.9157	0.9248	0.924	0.9440	0.9267	0.9311	0.9301	0.5398	0.5970
10−14	362	0.9376	0.9433	0.9421	0.9609	0.9457	0.9483	0.9481	0.5529	0.6117
Total	14,880	0.5842	0.6853	0.6638	0.6872	0.6646	0.6658	0.6733	0.5329	0.5948

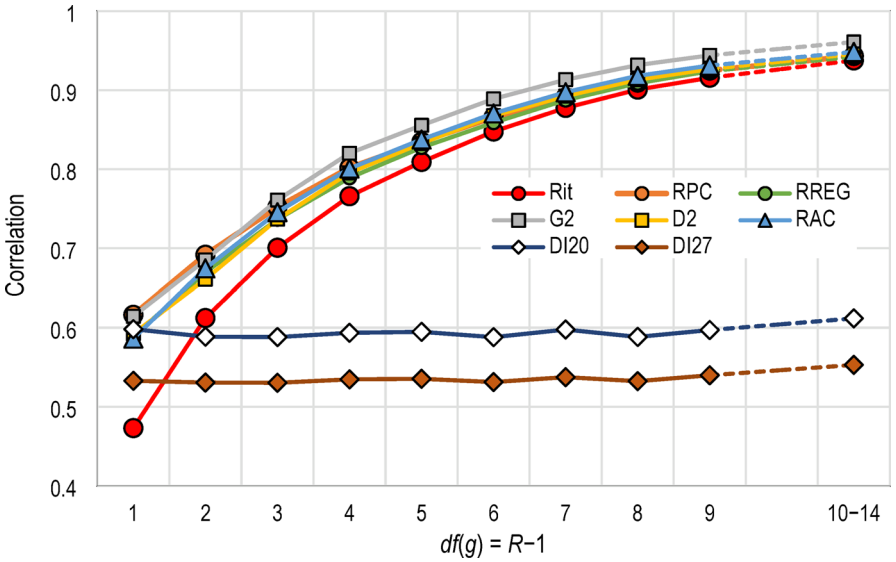

Fig. 3 Average estimates of item–score association by selected estimators by df(g)

Another feature of the traditional *DI* worth highlighting is that although it tends to give estimates slightly closer to the true association than *Rit* in the binary settings, it tends to radically and obviously underestimate the discriminative power with polytomous items. This can be inferred from its behaviour in relation to *Rit*, which tends to give lower values than *Rit* in the polytomous settings. Also, estimation by the traditional 27% cut-off seems to lead to an underestimation of item discrimination even for binary items when we compare *DI* with the better performing estimators. A better estimate by *DI* would be obtained by using the 20% cut-off or perhaps even the 16% cut-off (Fig. 4). Thus, contrary to what is suggested by Liu (2008) and Beuchert and Mendoza (1979), for example, the 30% or 33% cut-off for *DI* would move the esti-

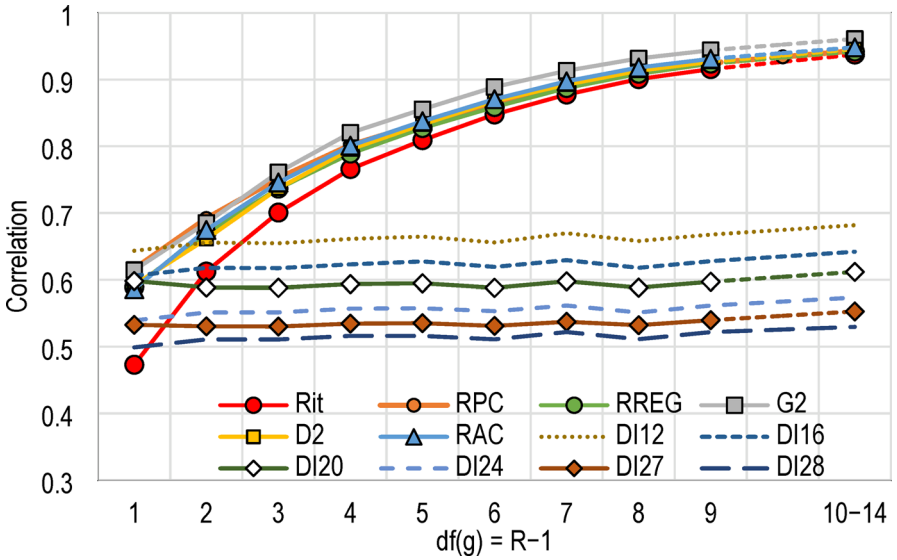

Fig. 4 Tendencies of the estimates by different cut-offs of DI

mate even further away from the true association, closer to a benchmark of *Rit* or perhaps even below and towards apparent underestimation, even for binary items. In the following, the 20% cut-off is taken as the basis for the dimensionally corrected *DI*.[5]

4.2 Dimension-corrected DI

In order to correct the illogical tendency of *DI* with polytomous items, a simple modification is proposed. The correction is based on the empirical observation that *DI* estimates tend to be "constant" regardless of the number of categories in the item. Therefore, the model related to $DI_{20\%}$ (model 0, M0) can be simplified as the rounded average of the estimates (see Table 1) in the following form:

M0: $DI_{20\%} \approx 0.59$.

However, it *should* follow the trend of the estimates of the better performing estimators. In the training dataset, the trends of the better-behaved estimators differ from each other to some extent, so their average tendency ("mean *RPC*+") is used as a benchmark for the correction (Fig. 5a and b).

The average trend can either be roughly modelled using a linear function (M1Lin) as follows:

$$\text{M1Lin} : 0.0465\text{df(g)} + 0.568 \tag{4}$$

[5] Since the problems with estimators such as *D* and *G* and the *DI* index are essentially due to their poor behaviour with respect to the number of categories, i.e., the dimensions of the tables, it seems sensible to shorten the "*DI* corrected for the number of categories" to "dimension-corrected *DI*". This mimics the names given to D_2 (dimension-corrected *D*; Metsämuuronen 2020c) and G_2 (dimension-corrected *G*; Metsämuuronen 2021a).

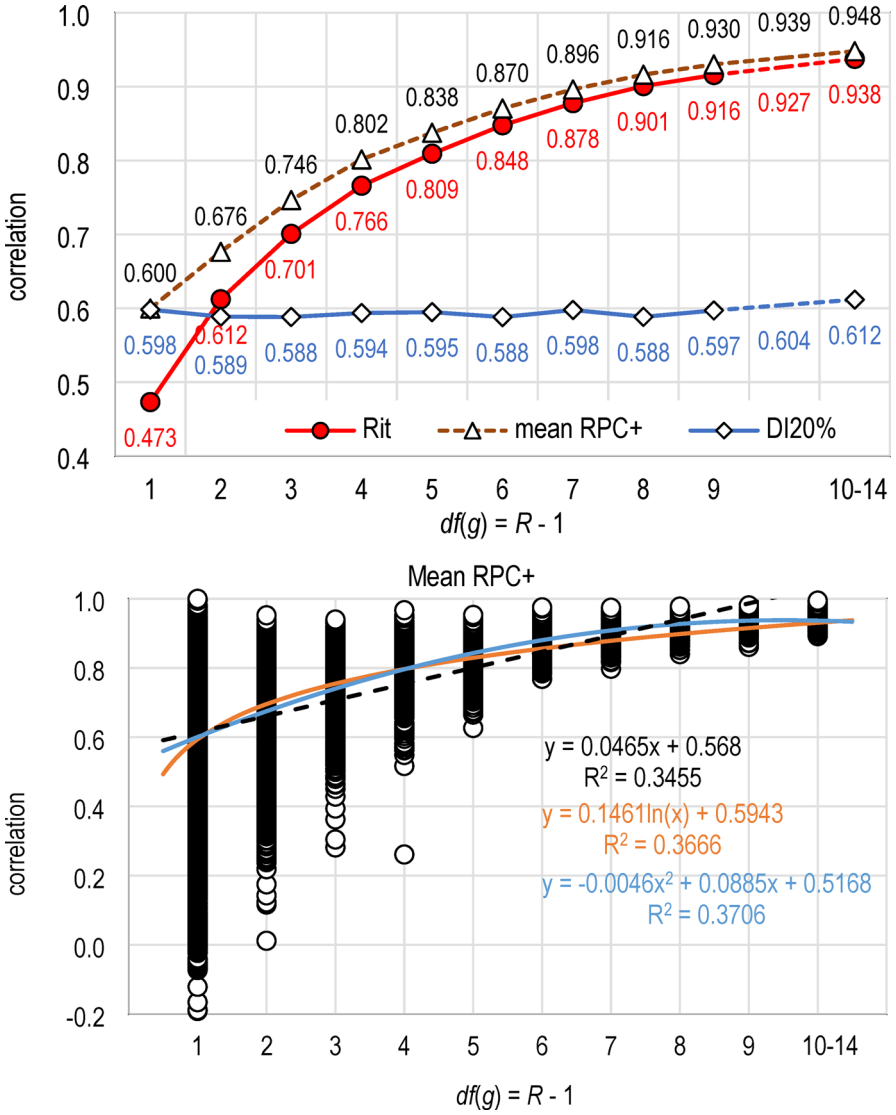

Fig. 5 a. Average estimates of $DI_{20\%}$ by df(g) and **b**. Mean estimates by the better performing estimators by df(g)

or, more precisely, by using a logarithmic function (M1Log) as follows:

$$\mathrm{MILog} : 0.1461\mathrm{LN}\left(\mathrm{df(g)}\right) + 0.5943 \tag{5}$$

or by a polynomial function (M1Pol)

$$\mathrm{M1Pol} \; - 0.0046(\mathrm{df(g)})^2 + 0.0885\mathrm{df(g)} + 0.5168 \tag{6}$$

All models fit the data points reasonably well in terms of $df(g)$ ($R^2 > 0.345$) and are rough approximations. Therefore, all may be rounded to some extent. Among the models, Eqs. (4) and (5) may be more interesting than Eq. (6) because by rounding them slightly, we get simple estimators without unnecessary constants or coefficients related to the training data set. After a convenient rounding, M1Lin could be as follows:

$$\mathrm{M1Lin}: \ 0.05\mathrm{df(g)} + 0.59 \tag{7}$$

and M1Log as follows:

$$\mathrm{M1Log} = 0.146LN\left(df(g)\right) + 0.59 \tag{8}$$

With these simplified models, the correction factor is the difference between the "constant" $DI_{20\%}$ (M0) and the expected level (M1Lin or M1Log):

$$\begin{aligned} \text{Linear Correction} &= \mathrm{M1Lin} - \mathrm{M0} \\ &= 0.05df(g) + 0.59 - 0.59 \\ &= 0.05df(g) \\ &= 0.05(R-1) \end{aligned} \tag{9}$$

$$\begin{aligned} \text{Logaritm Correction} &= \mathrm{M1Log} - \mathrm{M0} \\ &= 0.146LN\left(df(g)\right) + 0.59 - 0.59 \\ &= 0.146LN\left(df(g)\right) \\ &= 0.146LN(R-1) \end{aligned} \tag{10}$$

and

There is no specific reason to correct the estimate by $DI_{20\%}$ for binary items. Therefore, the correction factor should be 0 when $df(g)=1$. Since $LN(1)=0$, correction factor in Eq. (10) needs no further treatment. However, the linear option is modified to switch off the correction for binary items. This is done by changing it to the following form:

$$\text{Linear Correction} = 0.05\left(df(g) - 1\right) = 0.05\left(R-2\right) \tag{11}$$

There are then two options for the dimension-corrected DI (DI_2) :

$$DI_2Lin = DI_{20\%} + 0.05\left(df(g) - 1\right) = DI_{20\%} + 0.05\left(R-2\right) \tag{12}$$

and

$$DI_2Log = DI_{20\%} + 0.146LN\left(df(g)\right) = DI_{20\%} + 0.146LN\left(R-1\right) \tag{13}$$

In general, both corrections behave as intended: for binary items, $DI_2 = DI_{20\%}$, and for polytomous items, they raise the mean values by $DI_{20\%}$, close to or very close to the level of the better performing estimators (Fig. 6a and b). Among the options, the linear correction does not seem suitable for the items with more than 8 categories.

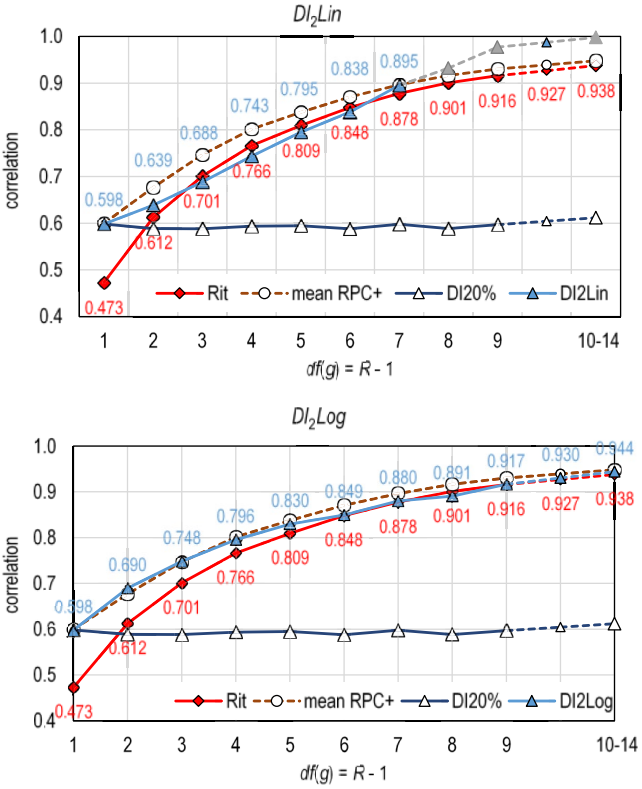

Fig. 6 a. The average estimates by the linear correction and **b**. The average estimates by the logarithm correction

It also tends to be very conservative for items with 4 to 7 categories; it tends to give estimates that are, on average, lower in magnitude than those by *Rit*. The logarithmic correction behaves quite optimally on the training data set: for items with 6 categories, it follows the trend of the better performing estimators, and for items with more than 6 categories, it tends to give conservative average estimates, following the trend of *Rit*. The characteristics of the logarithmic correction are examined below. However, it will be seen later on an external dataset that the more conservative linear model fits the Likert-type dataset better than the logarithmic correction. This justifies keeping the linear option as the recommended option as well.

4.3 Characteristics of the dimension-corrected DI

Since the *DI* itself is a crude, "quick and dirty" indicator of item discrimination or item validation to begin with, it may make sense that the correction should also be crude and simple enough to be easy to use in practical test settings. The suggested correction tends to bring estimates closer to the true association. However, under certain conditions, there is "collateral damage". As the intention was to keep the correction as simple and usable as possible from the practical user's point of view,

no complicated switches are proposed to switch off the correction when the value $DI_2 = DI_2Log = 1$ is reached, nor for the case that $DI_{20\%} = 0$. In particular, the formulae for G_2 and D_2 contain two such switches (see Metsämuuronen 2021a). Therefore, they do not exceed the value $G_2 = D_2 = 1$.

The average trends of *DI*, $DI_2 = DI_2Log$ and benchmarks for the lower limit (*Rit*) and the "true" association ("mean *RPC*+") are summarised in Table 3. The table will not be commented on here; the characteristics of DI_2 will be discussed later. First, the limitations and shortcomings of DI_2 are discussed. Second, its characteristics are examined in comparison with the benchmark estimators and with the theoretical "average" estimate of the better performing estimators. Third, the deflationary properties of DI_2 are discussed. Finally, the behaviour of the proposed dimension-corrected *DI* is challenged using the external datasets.

4.3.1 General characteristics, boundaries, and deficiencies of DI_2

The high or high-ish correlations between the estimates of the better performing association estimators and the traditional *DI* (0.585 for $DI_{27\%}$ and 0.543 for $DI_{20\%}$; see Table 4) confirm that *DI* captures something like the same phenomenon as correlation (*RPC*, *RREG*, *RAC*) and probability (G_2, D_2). This is a known property of *DI*—otherwise we would not use it. However, traditional *DI* estimates tend to be very conservative, especially for polytomous items. Compared to the traditional *DI*, DI_2Log correlates notably higher with both *Rit* (0.838) and the mean of the better performing estimators (0.785), as well as with the traditional *DI* (0.736–0.846). In parallel, DI_2Lin correlates notably higher with both *Rit* (0.801) and the mean of the better estimators (0.776), as well as with the traditional *DI* (0.753–0.868). The logarithmic correction changes the "constant" nature of the estimates by *DI* to be more logistic in nature (see Fig. 7 above). Even if we do not know the true item discrimination powers in the semi-simulated training dataset, it is noteworthy that the better performing estimators do not have deflation or the deflation is very small and, thus, they reflect the latent association better than *Rit* which is always deflated in the measurement modelling settings. Since the estimates of DI_2 are close to those estimates that are known (or believed) to be close to the true association, we may trust that DI_2Log and DI_2Lin are in a "good company" even without knowing what the exact discrimination would be. Of course, systematic studies would be beneficial in verifying this.

Looking at the overall distributions of $DI_{20\%}$ and DI_2Log, two characteristics are worth highlighting. First, they both share a tendency to underestimate when the item difficulty is extreme (see also Tristan 1998); this characteristic is highlighted in Fig. 8 as grey areas in the scatter plots related to $DI_{20\%}$ and DI_2Log. Therefore, for those items with extreme difficulty level ($0.20 > p > 0.80$), it is advisable to use a different estimator of the item-score association; any of the better performing estimators (*RPC*, *RREG*, *G*, G_2, *D*, D_2, *RAC*, or *EAC*) would be a good option. Second, although DI_2 effectively raises the estimates by $DI_{20\%}$ to the same level as the better performing estimators in the polytomous cases, the variability of the estimates with a very large scale appears to be greater with DI_2Log compared to the benchmark estimators. This was seen in Table 3 with slightly higher standard deviations and seems to be inherited from the mechanics of calculating the *DI*.

Table 3 Average estimates by DI_2 and selected benchmarking estimators

df(g)	*N*	Mean					Std. Deviation				
		$DI_{20\%}$	$DI_{27\%}$	DI_2=DI_2Log	Mean[1]*RPC*+	*Rit*	$DI_{20\%}$	$DI_{27\%}$	DI_2=DI_2Log	Mean[1]*RPC*+	*Rit*
1	7,948	0.598	0.533	0.598	0.600	0.473	0.206	0.186	0.206	0.159	0.136
2	3,056	0.589	0.531	0.690	0.676	0.612	0.143	0.134	0.143	0.109	0.110
3	1,390	0.588	0.530	0.748	0.746	0.701	0.113	0.107	0.112	0.083	0.085
4	729	0.594	0.535	0.796	0.801	0.766	0.101	0.097	0.100	0.066	0.071
5	474	0.595	0.535	0.830	0.838	0.809	0.090	0.086	0.090	0.052	0.058
6	366	0.588	0.531	0.849	0.870	0.848	0.073	0.071	0.073	0.034	0.037
7	255	0.598	0.537	0.880	0.896	0.878	0.082	0.080	0.079	0.031	0.032
8	140	0.588	0.532	0.891	0.916	0.901	0.072	0.068	0.070	0.026	0.028
9	160	0.597	0.540	0.917	0.930	0.916	0.058	0.057	0.056	0.024	0.026
10–14	362	0.612	0.553	0.944	0.948	0.938	0.050	0.046	0.044	0.015	0.017
Total	14,880	0.595	0.533	0.674	0.673	0.584	0.171	0.156	0.196	0.164	0.182

1) mean of the estimates by RPC, RREG, G2, D2, and RAC

Since the aim was to provide a simple tool for practical use, and since *DI* itself is a crude tool for investigating item discrimination or validating items, DI_2Lin and DI_2Log are also crude indices. This can be seen in two shortcomings related to polytomous items, which are not taken into account in the proposed simple formula. First, for polytomous items, DI_2Lin and DI_2Log estimates will automatically exceed 1 if the $DI_{20\%}$ estimate is very high to begin with. This can be strictly predicted from Eqs. (12) and (13); for $df(g)=2$, the values $DI_{20\%}>0.890$ lead to values $DI_2Log>1$, and with $DI_{20\%}>0.950$ lead to values $DI_2Lin>1$. Correspondingly, in case of DI_2Log, the thresholds for $df(g)=3$, 4, 5, 6, 7 and 8 are 0.826, 0.781, 0.746, 0.717, 0.693 and 0.671 respectively. However, it appears that DI_2Log is safe to use up to 7–8 categories in the item; with 5–7 categories in the item, the prevalence of out-of-range values in the training data set is only 2.5%. The prevalence increases with the number of categories: with 8 to 10 categories the prevalence is 12.3%, and with more than 10 categories the prevalence in the training data set is 31%; out of 362 estimates with $df(g)=10$–14, 112 were found to be higher than 1. These figures obviously depend on the original data set used in the simulation. However, they may give a rough approximation of the prevalence for the out-of-range estimates.

The highest estimate for DI_2Log obtained in the training data set is 1.103 and for DI_2Lin, 1.217; the corresponding average estimate of the better estimators is 0.989; most of the estimates outside the range of DI_2 seem to be close to the value 1, if not exactly 1 by the better estimators. Therefore, a simple solution to this deficiency is proposed: the values exceeding $DI_2=1$ are truncated to 1. Although this is not an optimal solution, its logic corresponds to the practicality of correcting the biased estimates by eta squared with the unbiased alternatives omega squared and epsilon squared. The latter tend to give negative outlier estimates of the explanatory power when η^2 is close to 0, especially with the small sample sizes (see, for example, the discussion in Okada 2013, 2017). For these outliers, the negative results are traditionally replaced by 0. In Fig. 6b above, the out-of-range estimates are truncated to 1. This has no visible effect on the average of the estimates.

Another source of roughness is that for polytomous items, when $DI_{20\%}=0$, DI_2 is always greater than 0. However, this is a very rare case in the training data set; only two such cases on this kind were found. However, strictly from Eq. (13) it is known that when $DI_{20\%}=0$ and $df(g)=2$, 3, 4, 5, 6, 7, or 8, the values by DI_2Log are 0.110, 0.174, 0.219, 0.254, 0.283, 0.307, and 0.329, respectively. As a shortcoming, this mechanical or technical increase in the magnitude of the non-existent association may be a more serious one than the out-of-range estimates discussed above. For the strict value of $DI_{20\%}=0$ we can use a simple rule to cut DI_2 to 0 although the prevalence of this is low. However, the values $DI_{20\%}=0$–0.20 are more challenging. If we take seriously Ebel's (1965) traditional rule of thumb that a value of $DI<0.20$ indicates a poorly functioning item that needs to be omitted or completely revised, we may be willing to reconsider this boundary for DI_2 with items with wide scale. Possible specific boundaries for DI_2 with polytomous items can be derived from the thresholds for $DI_{20\%}=0$. With polytomous items, we can convert the traditional boundary for the lowest acceptable level of the estimate by DI_2Log: with $df(g)=2$, 3, 4, 5, 6, 7, or 8, the lower boundary could be 0.31 ($=0.101+0.20$), 0.37, 0.42, 0.45, 0.48, 0.51, and 0.53, respectively. For example, for a 5-point Likert-scale item or a test item with a score of

Table 4 Correlations of the estimates for selected estimators of association, DI, and DI_2

n=14,880	DI20%	DI27%	DI2Log	DI2Lin	Mean RPC+	Rit
DI20%	1	0.859	0.864	0.868	0.543	0.503
DI27%		1	0.751	0.753	0.585	0.545
DI2Log			1	0.988	0.785	0.838
DI2Lin				1	0.776	0.801
Mean RPC+					1	0.945
Rit						1

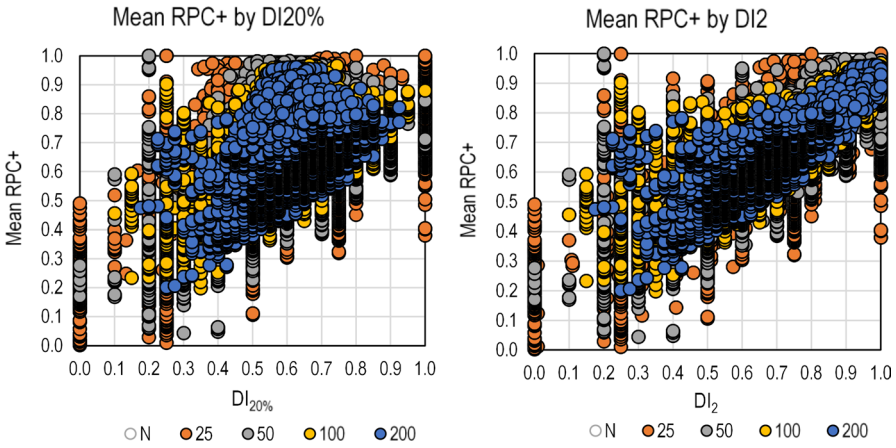

Fig. 7 Distributions of the estimates by $DI_{20\%}$ and DI_2Log with the better performing estimators

0–4 with $df(g)=4$, we can conclude that an item with $DI_2Log<0.419$ (=0.219+0.2) would not be discriminatory enough to be selected for the test because the estimate contains a significant amount of artificial increase in the estimate.

Another source of roughness in DI_2, inherited from the computational process of the traditional DI, is that, unlike the better performing benchmarking estimators, it cannot detect perfect discrimination with polytomous items unless all cases in the lower part receive the lowest possible item score and all cases in the upper part receive the highest possible item score. The benchmarking estimators do not depend on the "lowest possible" and "highest possible" options. In binary cases this is not an issue, but in polytomous cases both DI_2 and DI are vulnerable in this respect.

In terms of sample sizes and the number of categories in the items, it is clear that DI_2 gives slightly more spread out estimates with large scale items compared to the benchmark estimators (Fig. 9). This is inherited from the behaviour of DI. However, this phenomenon is reversed for items with a narrow scale; up to 4 categories the variability seems to be lower with DI_2 than with the benchmarking estimators, and above 5 categories the variability seems to be higher with DI_2 than with the benchmarking estimators.

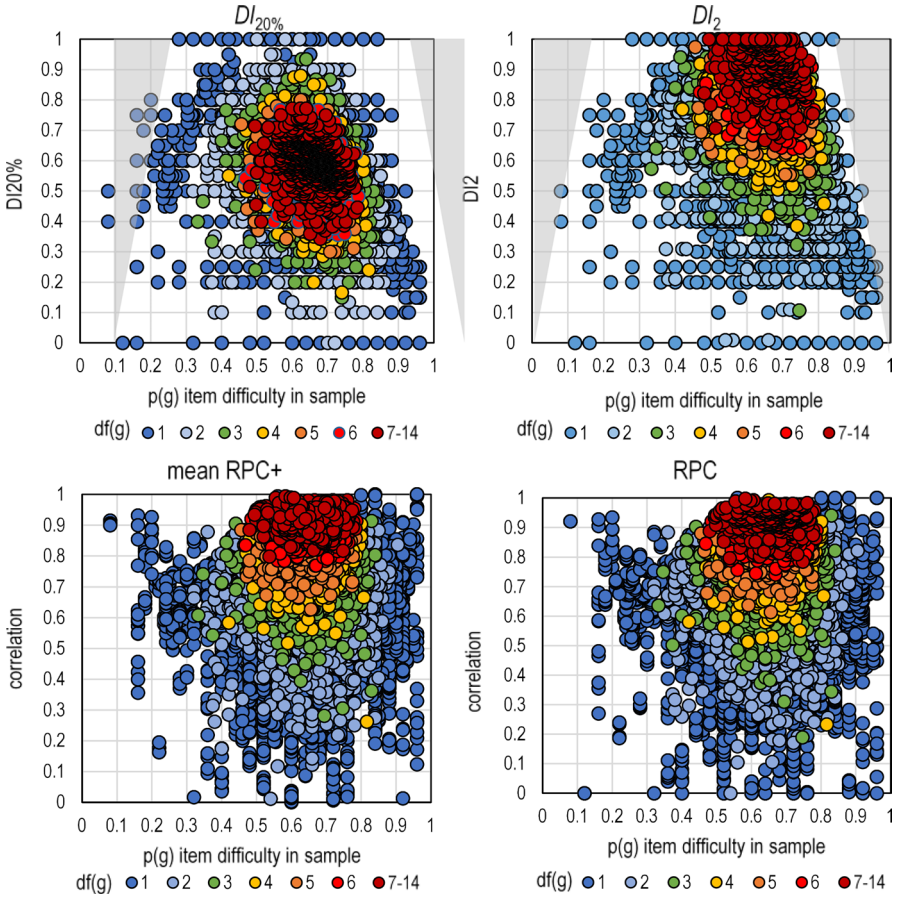

Fig. 8 General distributions of $DI_{20\%}$, DI_2Log, RAC, and the mean of the better performing estimators

4.3.2 Tendencies of DI_2Log under selected conditions

The tendencies in DI_2Log are examined in more detail in comparison with relevant indicators of the test and the data set. The distributions of $DI_{20\%}$ and DI_2Log estimates are already compared by item difficulty (see Figs. 7 and 8 and related discussion) and item scale (see Fig. 9), and some effect of sample size can be seen in Figs. 7 and 9 above. Here their tendencies are compared under four conditions relevant to testing settings: sample size, item difficulty, number of categories in the score, and number of items in the score. Figure 10 summarises the results.

In terms of sample size (N), the *DI20%* and *DI2* estimates tend to produce slightly higher estimates than the benchmark estimates, although the difference is not remarkable—around 0.02 units of correlation on average. Figures 7 and 9 above show that the variability of the estimates tends to decrease with sample size.

In terms of item difficulty (p), the differences between the estimators are striking. Both $DI_{20\%}$ and DI_2 tend to underestimate items with extreme difficulty levels, as

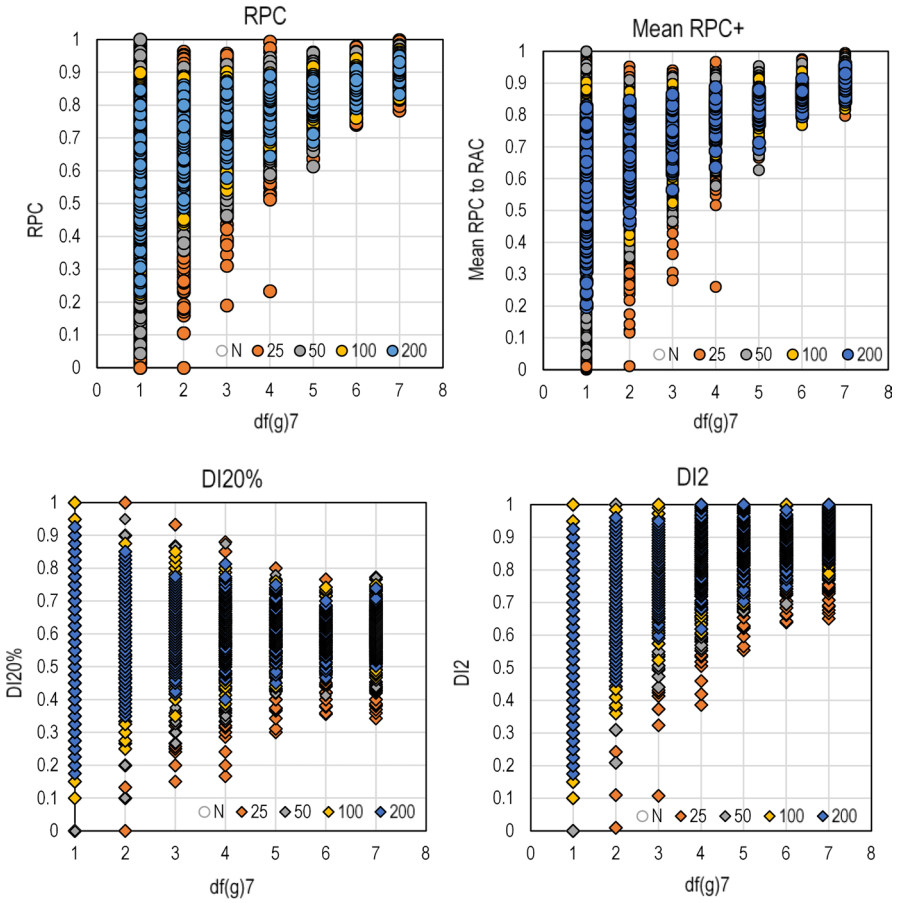

Fig. 9 Distributions of $DI_{20\%}$, DI_2, RAC, and the mean of the better performing estimators in the training dataset

discussed above. From this point of view, the better performing estimators give more stable estimates. For items of medium difficulty, the DI_2 estimates tend to be higher than the other estimators. This is partly due to the characteristic of the traditional DI to detect the deterministic pattern by the value $DI=1$ and partly due to the mechanical increase for these estimates. It may be too early to say that the estimates are overestimates—this would require more systematic studies. However, this phenomenon seems to be related to polytomous items.

In terms of the number of categories in the score, as indicated by $df(X)$, the estimates seem to be stable when the test score has more than 15 categories. Below this, the estimates tend to increase with the number of categories. This is also the tendency for the benchmarking estimates.

In terms of the number of items in the test (k), it is good to note an obvious confounding factor. The training data set is formed so that the wider the scale in the items is, the fewer items there are in the test. Therefore, the number of items in the test cannot be separated from the number of categories in the items. In Fig. 10, $k=20$–30

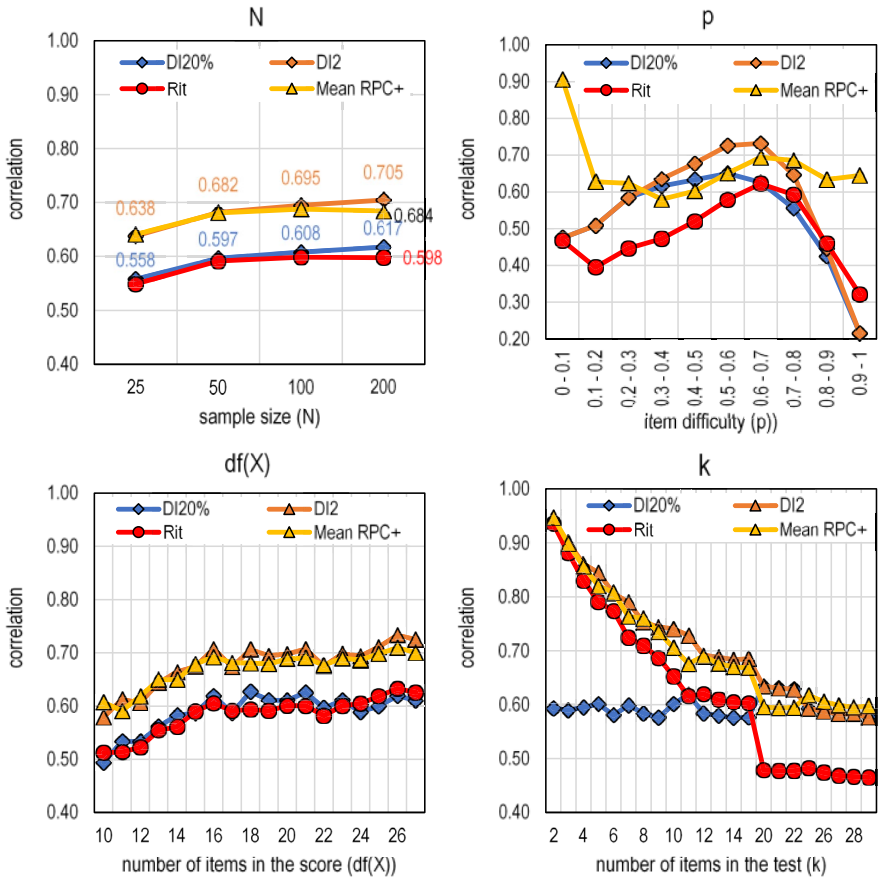

Fig. 10 Comparison of the tendencies of the estimators (DI_2=DI_2Log)

refers to binary items ($df(g)=1$), $k=11$–15 refers to items with $df(g)=2$, $k=7$–10 refers mainly to items with $df(g)=3$, and $k=2$–3 refers to items with $df(g)=10$–14. Nevertheless, $DI_{20\%}$ behaves notably differently from the other estimators, as it tends to give "constant" estimates regardless of the number of items in the test or the number of categories in the items; this was the reason for proposing the modification. It also appears that, with binary items, DI_2 (=$DI_{20\%}$) is more stable than *Rit* or even the better performing estimators in general.

4.3.3 Behaviour of DI_2Log in the deterministically discriminating settings

The better performing estimators of item score association are "better" because the estimates do not include mechanical or artificial deflation when two identical variables are correlated in different way—or the amount of deflation is nominal. That is, unlike *Rit*, for example, they reflect the latent association error-free or close to it, regardless of the item difficulty. We do not know how *DI* or DI_2 behave in this regard.

Hence, the simulation design of Metsämuuronen (2021a), (2022c) is replicated to investigate this.

In the design, three vectors are formed with $N = 1000$ cases: a standardised normal vector with N(0,1), a uniform vector with no tied cases, and a skewed normal vector with Γ(2,1). Each vector is duplicated to form three pairs of (perfectly correlated) variables. Each pair of vectors is manipulated so that one of the identical vectors becomes a variable with a wider scale (score X) and the other with a narrower scale (item g). By changing the cut-offs of the original vector, the scale of X is set to vary with respect to the normal and gamma distributions 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 scale of g is set to vary with fixed values $df(g) = 1$, 2, 3, and 4, i.e., the most commonly used scales from the binary scale to a 5-point Likert type or 0–4 point scale. The item difficulty is varied by systematically changing the cut-off for the bins. The construction of the dataset is described in detail in Metsämuuronen (2021a). The final dataset contains 22,824 estimates from each estimator of interest. The dataset is available in CSV format at https://doi.org/10.13140/RG.2.2.14598.45127 and in SPSS format at https://doi.org/10.13140/RG.2.2.31375.66726. Figure 11 summarises the main results. Two points are highlighted from the analysis.

First, the traditional *DI* underestimates obviously the perfect association with polytomous cases. This is partly understandable because of computational mechanics discussed above. While the better performing estimators detect the deterministic pattern regardless of the maximum value of the item, the magnitude of the estimates by *DI* depends strictly on the distance between the values obtained in the upper and lower parts of the dataset. With a normal distribution, both $DI_{27\%}$ and DI_2Log tend to include the more deflation the higher the number of categories in the item gets. In particular, neither $DI_{27\%}$ or DI_2 reach the perfect value $DI_{27\%} = DI_2Log = 1$ when four categories are reached. With 5 categories in the item, even in the best case,

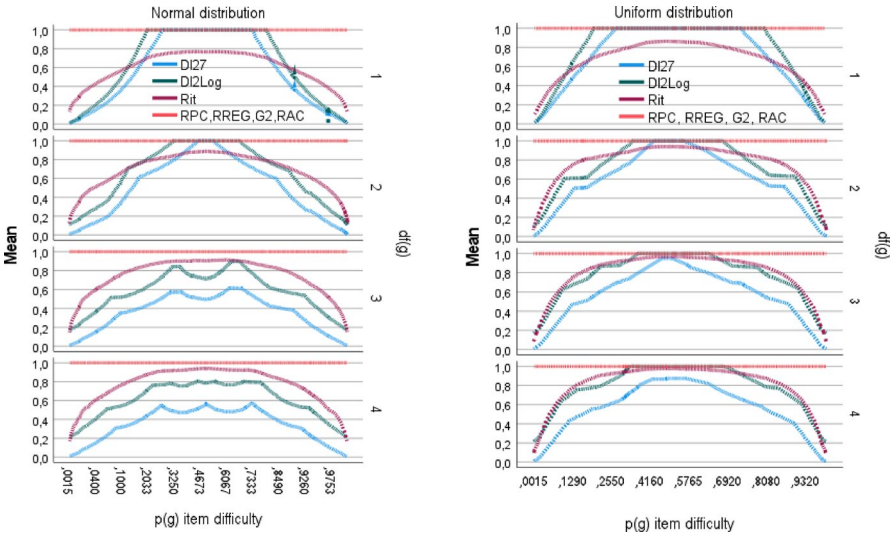

Fig. 11 Deflation in DI_2Log and in benchmarking estimators

$DI_{27\%}$ contains notably more deflation (>40%) than *Rit* (>5%) and notably more with items with extreme difficulty levels. On the other hand, with a uniform distribution, DI_2Log, unlike *Rit*, reaches the perfect (true) value $DI_2Log = 1$ with the mediocre items. Otherwise, it behaves in a similar way to *Rit*: the deflation is remarkable for items with extreme difficulty level.

Secondly, both $DI_{27\%}$ and DI_2Log behave better with the uniform distribution than with the normal distribution. The fact that the behaviour of DI_2Log with uniform distribution is decent can be seen as an advance in practical settings with (presumably) small sample sizes. Compared to the large sample size settings, in the small sample size settings it is more likely that all test takers will receive a unique score, i.e., the test score can separate all test takers from each other. The score is then automatically uniformly distributed and DI_2Log behaves as optimally as possible given its tendency to produce radical underestimates on items of extreme difficulty. Although both $DI_{27\%}$% and DI_2Log perform better with uniformly than with normally distributed scores, with polytomous items DI_2Log performs significantly better than $DI_{27\%}$% even in the normally distributed case, although not much better than *Rit*. In this respect, both $DI_{27\%}$% and DI_2Log belong to the group of "worse performing" estimators rather than to the group of "better performing" estimators.

4.3.4 How DI_2 behaves in the hypothetical dataset in Table 1

A reader may be interested to know how DI_2 behaves in the hypothetical, deterministically discriminating data set discussed with Table 1. The traditional $DI_{27\%}$ with four cases in both extremes gave an estimate of 0.750 for both items while the better performing estimators detected the latent perfect association ($RREG \approx RPC \approx G_2 = D_2 = RAC = 1$), and the estimate by *Rit* was notably lower (0.664). In the hypothetical data set with three cases at both extremes, $DI_{20\%} = DI_2Log = DI_2Lin = 1$ for the binary item and, with the polytomous item with $df(g) = 4$, $DI_{27\%} = 1–2/12 = 0.833$ giving $DI_2Log = 0.833 + 0.146 \times LN(4) = 1.036$ which is cut to $DI_2Log = 1.000$, and $DI_2Lin = 0.833 + 0.05 \times 3 = 0.983$, which is a slight underestimate. It seems that DI_2 behaves expectedly well in the given hypothetical, deterministic dataset: its values correspond with the estimates by the better performing estimators of association.

4.3.5 How DI_2 behaves outside on the training dataset

As DI_2 was developed using a semi-simulated training dataset, it is interesting to see how it performs in independent real data sets. Clearly, more extensive simulations are needed to investigate and confirm its behaviour in controlled settings with different factors and different true item discrimination values. Here are two examples of the behaviour of the modified *DI* with an external data set. The first example relates to a 5-point Likert scale attitude test and the second to an independent achievement test with binary and polytomous items.

The first data set consists of 15 attitude items with a 5-point Likert scale. The test is a modified version of the Fennema-Sherman Attitude Scales (Fennema and Sherman 1976; see validation in Metsämuuronen 2012). A shortened version of the original test includes three dimensions: "liking mathematics", "self-efficacy in

mathematics" and "usefulness in mathematics", each with five items. In each dimension, one or two items are reversed. Only the total score is used in the analysis. For the analysis, all items are treated in a positive direction. The reliability of the total score is high or highish in the datasets related to higher grades in compulsory education (α = 0.86–0.92; Metsämuuronen 2012, 2023; Metsämuuronen and Nousiainen 2021; Metsämuuronen and Tuohilampi 2014; deflation-corrected $\omega_{D2} = 0.94$; Metsämuuronen & Nousiainen 2023; for deflation-corrected reliability estimates, see Metsämuuronen 2022i). The dataset itself is not open, but the results on attitudes are reported in Metsämuuronen (2023) and the structure of the test is examined in Metsämuuronen (2012) and Metsämuuronen and Nousiainen (2023). Here, the data set is re-analysed using the relevant item discrimination indices, *RAC*, D_2 and G_2. Their estimates are close to each other in the same way as in Figs. 3 and 4, and to condense the information their mean ("mean RAC+") is used as a reference. The dataset includes 11,235 test takers in a national mathematics achievement test of grade 9 students.

Figure 12 illustrates the result using a Likert scale. Four observations can be made. First, the traditional $DI_{27\%}$ index radically underestimates item discrimination in all items, as does the proposed $DI_{20\%}$. Their values are significantly lower than *Rit*, which is known to underestimate the association in all cases related to measurement modelling due to deflation. Taking item 7 as an example, while *Rit* = 0.649, $DI_{20\%}$ = 0.510 and $DI_{27\%}$ = 0.446. Secondly, the better performing estimators ("mean *RAC*+", the dotted line in Fig. 12), which are assumed to be closer to the true association due to less or no deflation, give a consistent message of item discrimination. Their tendencies overlap and are always higher than the tendency of *Rit*, which is too low due to deflation. Thirdly, in the case of the logarithmic correction to $DI_{20\%}$, "DI_2Log" tends to give estimates that are higher than the better performing estimators, i.e., it seems to overestimate item discrimination. Notably, the more conservative linear

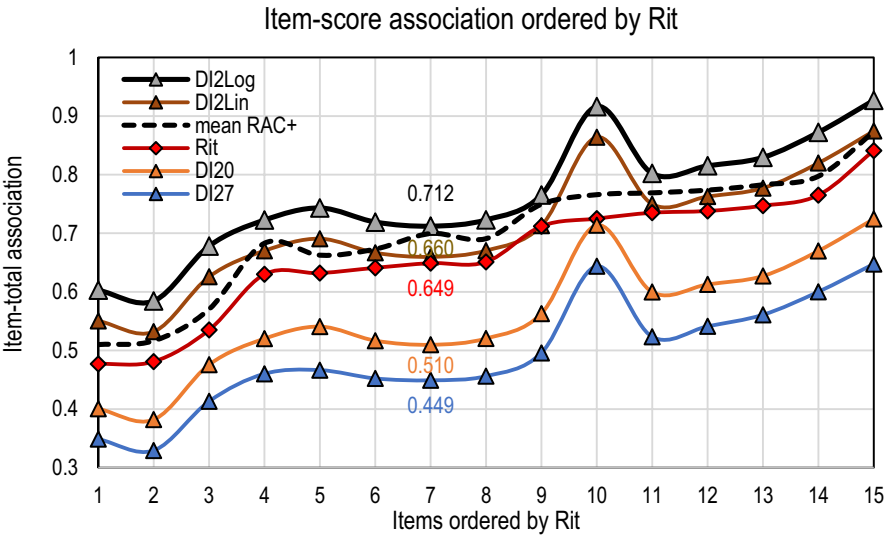

Fig. 12 Item discrimination in attitude test by selected indices

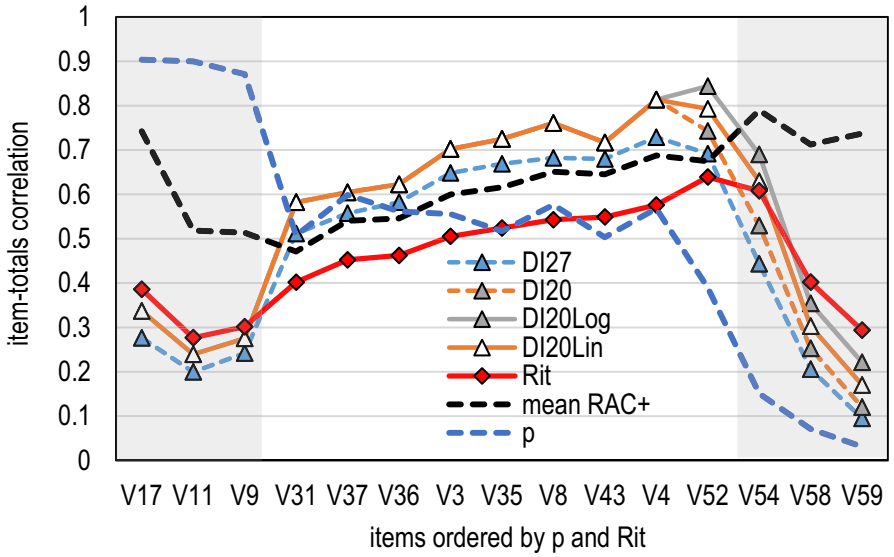

Fig. 13 Item discrimination in achievement test by selected indices

correction ("DI_2Lin") is closer to the trend of the better performing estimators. Thus, even though the estimates for DI_2Lin are conservative in the training dataset, they seem to be closer to the "true association" than those of $DI_{27\%}$ and $DI_{20\%}$, and even DI_2Log in the external data set. In item 7, the estimate for $DI_2Log = 0.712$, which uncharacteristically is close to the mean estimate of the better performing estimators, and $DI_2Lin = 0.660$, which seems to be a slight underestimate. Fourthly, all variants of *DI* detect a distinctly high discrimination in item 10. For this particular item, both DI_2Log and DI_2Lin seem to give an overestimate, and this is caused by an unexpectedly high estimate from $DI_{20\%}$. Overall, DI_2Log and DI_2Lin seem to behave as expected with the attitude scale outside the training data set: the estimates give a higher estimate than $DI_{27\%}$ and $DI_{20\%}$, they follow the better performing estimators. Of these, DI_2Log may give slight overestimates with 5-point Likert scale.

The second example comes from a mathematics achievement test for grade 9 students (Metsämuuronen 2023). One of the seven versions of the test with 60 items ($n = 4{,}451$) is re-analysed from the perspective of *DI*. The reliability of the total score is high (traditional $\alpha = 0.922$, $\omega = 0.927$, deflation-corrected $\alpha_D = 0.960$, $\omega_D = 0.966$; Metsämuuronen & Nousiainen 2023). Notably, only four of the items are polytomous, and three of them were very difficult ($p < 0.15$), so that *DI* is expected to produce radical underestimates (see discussion with Fig. 8). To condense the example, only the three easiest items ($p > 0.871$) and the three most difficult polytomous items ($p < 0.15$) are shown in Fig. 13, and the number of items of medium difficulty is reduces to nine example items. The items are first ordered by proportion of correct answers (p), then the medium items are ordered by *Rit*. The difficulty levels at which *DI* is expected to produce radical underestimates are shaded in grey.

Three remarks are made. First, except the for the four last items in Fig. 13 (V52, V54, V58, V59), all items are dichotomous, i.e., $df = R - 1 = 1$. Therefore,

DI_2Log and DI_2Lin, which are based on $DI_{20\%}$, follow the same trend; with $df=1$, $DI_2Log=DI_2Lin=DI_{20\%}$. Secondly, Fig. 13 illustrates how *Rit*, $DI_{27\%}$, $DI_{20\%}$ as well as the DI_2Log and DI_2Lin underestimate the item discrimination power when the item difficulty is extreme. In these cases, the better performing estimators, with no or notably less deflation, capture the property of the items to be highly discriminative. Third, in the external dataset, the traditional $DI_{27\%}$ tends to give estimates for the items of medium difficulty levels that closely follow the tendency of the better performing estimators. For these items, the $DI_{20\%}$ and, consequently, the DI_2Log and DI_2Lin tend to overestimate, except in the range where the traditional *DI* gives underestimates. In this range, for polytomous items, the DI_2Log gives the closest estimate to the better performing estimators, but it is still much too low.

Two general points can be made using the external dataset. First, even though the $DI_{20\%}$ gives estimates closer the better performing estimators than the $DI_{27\%}$ with the binary items in the training dataset, the external dataset with binary items challenges this pattern. Second, both the $DI_{20\%}$ and $DI_{27\%}$ seems to give notable underestimation with Likert type scales in both the external and training datasets. Both DI_2Log and DI_2Lin give estimates that are, on average, close the better performing estimators in the training dataset. However, an external dataset with polytomous Likert items challenges this pattern. It appears that DI_2Log may overestimate item discrimination power with the Likert-type of scale, while the more conservative estimator DI_2Lin may give estimates closer to the better performing estimators. Both examples with external datasets suggest that there is a need for systematic simulations in relation to the proposed corrections for *DI* in both the binary and polytomous settings.

5 Discussion

5.1 Results in nutshell

The starting point for the proposal of a dimension-corrected discrimination index or upper-lower index was twofold. First, the traditional 27% cut-off *DI* appeared to underestimate item discrimination. For binary items, the magnitude of the estimates from $DI_{27\%}$ tended to be notably lower than those from the better performing estimators such as *RPC*, *RREG*, G_2, D_2, and *RAC* although they were quite close to the estimates from the point-biserial correlation. However, for polytomous items, the $DI_{27\%}$ estimates tended to be obviously underestimated. For binary items, $DI_{20\%}$ estimates tend to be comparable to the average estimates of the better performing estimators. On the other hand, *DI* behaves illogically compared to the benchmarking better performing estimators in the sense that while the better performing estimators based on correlation and probability (correctly) approach the value $\rho=1$ as the number of categories in items approximates the number of categories in the score, the estimates by *DI* tend to be "constant" and obviously too low in magnitude regardless of the number of categories in the item.

Two dimension-corrected *DI*s have been proposed, the logistic modification $DI_2 = DI_{20\%} + 0.146LN\,(R-1)$ and the linear modification $DI_2 = DI_{20\%} + 0.05\,(R-2)$, where *R* is the number of categories in the item. Both

behave as intended on the whole: for binary items, DI_2 equals $DI_{20\%}$, and for polytomous items, DI_2 raises the estimates by $DI_{20\%}$ to the same level as the better performing correlation estimators. In addition, neither includes unnecessary additional constants or coefficients related to the training data set that a polynomial correction included. Also, both proposals for DI_2 perform notably better than $DI_{27\%}$ in terms of deflation in the estimates, especially for uniformly distributed scores. This may be an expected form of score distribution in the practical test settings with small sample sizes.

However, since the aim was to develop a simple tool for the practical user, the dimensional correction does not include specific switches for censoring the out-of-range values; these may be obtained with the polytomous items if the original uncorrected $DI_{20\%}$ initially obtains a high value. For these cases, it is suggested that whenever an out-of-range estimate is obtained, it should be trimmed to $DI_2 = 1$ assuming that the item is not ultimately pathological in nature and exceeds the *negative* limit (-1). The prevalence of such values is quite high when the number of categories exceeds 8 or 9. However, this type of item with such a large scale is very rare in real-life testing situations, except in essay-type tests. For these items, *DI* and DI_2 are not the best options.

Another challenge with dimension-corrected *DI* is that, for polytomous items, the correction gives an artificial boost to estimates with very low and non-existent magnitude. The more categories in the item, the more artificial increase is embedded in the estimates. This artificial increase is proportionally greater for items with low magnitude in the *DI*. Therefore, it is suggested that for polytomous items, the traditional lower bound for acceptable item discrimination ($DI > 0.20$) may need to be modified to depend on the number of categories in the items.

In addition, the dimensional correction does not correct the fundamental flaw in $DI_{20\%}$: items of extreme difficulty can be radically underestimated—even more so than by *Rit*. This characteristic of DI_2 is inherited from the traditional *DI*. For items of extreme difficulty levels, neither *DI* nor DI_2 is the best option to use; any of the better performing estimators such as *RPC*, *RREG*, *G*, G_2, *D*, D_2, *RAC*, or *EAC* would be a better alternative to *DI* or DI_2.

The external datasets used as practical examples challenge the results and suggest systematic simulations in relation to the proposed corrections for *DI* in both the binary and polytomous settings. First, in the training dataset, $DI_{20\%}$ gives estimates closer to the better performing estimators than $DI_{27\%}$. However, in the external dataset with binary items, $DI_{27\%}$ tends to give estimates closer to the better performing estimators while $DI_{20\%}$ tends to overestimate the item discrimination notably, particularly in the range of $0.27 < p < 0.75$. Second, in the polytomous settings, both the DI_2Log and DI_2Lin resolve the apparent and notable underestimation by $DI_{20\%}$ and $DI_{27\%}$. However, in the external dataset DI_2Log gave overestimates with the Likert-type of scale while the more conservative estimator DI_2Lin gave closer estimates to the better performing estimators.

Considering these shortcomings as part of the roughness of the Discrimination Index—after all, the *DI* is a rough tool to start with—we can say that the DI_2 (either DI_2Log or DI_2Lin) seems to be safe to use in the difficulty range $p = 0.20$–0.80 and when the number of categories in the item does not exceed 7–8. Items with these such

scales cover the most commonly used test settings. Systematic studies are needed to validate their behaviour.

5.2 Condensed suggestions for the practical settings

As the *DI* is a useful tool for item validation in the practical settings of testing when there is no need for precise estimates, the proposed modification, the dimension-corrected *DI*, improves the estimation, especially for the polytomous items. The estimates are still rough, as *DI* is a rough estimate to begin with, but they are most likely closer to the true association than when using the traditional *DI*. For the practical settings where *DI* is used, the following suggestions are summarised:

1) Instead of the traditional 27% cut-off, use a 20% cut-off; in the training dataset, this tended to give an estimate of the item discrimination power that is closer to the better performing estimators of item-score association than the traditional 27% or 30% cut-offs would give. If the traditional 27% cut-off is used, a constant of 0.06 could be added to the estimate; this tends to increase the estimate at the level of $DI_{20\%}$ in the training dataset.
2) When polytomous items are used, the following corrections based on the logarithmic correction $DI_2Log = DI_{20\%} + 0.146LN\left(df(g)\right) = DI_{20\%} + 0.146LN\left(R-1\right)$ could be used. This correction for the traditional *DI* tends to produce estimates that match the estimates of the better performing estimators in the training data set:

Number of categories in the item (*R*)	df(g) = *R*–1	0.146×LN(*R*–1)	DI_2 by using $DI_{20\%}$	DI_2 by using $DI_{27\%}$	Lowest value for acceptable item discrimination (by the rule of > 0.20)
2 (binary)	1	0	DI20%	DI27% + 0.06	0.20
3	2	0.10	DI20% + 0.10	DI27% + 0.16	0.30
4	3	0.16	DI20% + 0.16	DI27% + 0.22	0.36
5	4	0.20	DI20% + 0.20	DI27% + 0.26	0.40
6	5	0.23	DI20% + 0.23	DI27% + 0.30	0.43
7	6	0.26	DI20% + 0.26	DI27% + 0.32	0.46
8	7	0.28	DI20% + 0.28	DI27% + 0.34	0.48

3) Alternatively, the linear correction $DI_2Lin = DI_{20\%} + 0.05\left(df(g)-1\right) = DI_{20\%} + 0.05\left(R-2\right)$ could be used as follows:

Number of categories in the item (*R*)	df(g) −1 = *R*–2	0.05×(*R*–2)	DI_2 by using $DI_{20\%}$	DI_2 by using $DI_{27\%}$	Lowest value for acceptable item discrimination (by the rule of > 0.20)
2 (binary)	0	0	DI20%	DI27% + 0.06	0.20
3	1	0.05	DI20% + 0.05	DI27% + 0.11	0.25

Number of categories in the item (R)	df(g) $-1=R-2$	$0.05\times(R-2)$	DI_2 by using $DI_{20\%}$	DI_2 by using $DI_{27\%}$	Lowest value for acceptable item discrimination (by the rule of >0.20)
4	2	0.10	DI20% + 0.10	DI27% + 0.16	0.30
5	3	0.15	DI20% + 0.15	DI27% + 0.21	0.35
6	4	0.20	DI20% + 0.20	DI27% + 0.26	0.40
7	5	0.25	DI20% + 0.25	DI27% + 0.31	0.45
8	6	0.30	DI20% + 0.30	DI27% + 0.36	0.50

4) When either of the above corrections is used, estimates exceeding $DI_2=1$ are truncated to 1. These out-of-range values may occur when the number of categories is high and the value of the traditional DI is high to begin with. If the traditional $DI=0$, DI_2 can be truncated to $DI_2=0$.
5) If the traditional DI is very low ($DI=0.01$–0.20) for polytomous items, consider raising the lower limits for the acceptable item as shown in the tables above; the more categories in the item, the more artificial the increase in the estimate.
6) For extremely difficult items ($0.20>p>0.80$), it is better to use an estimator other than DI or DI_2.

5.3 Known restrictions and suggestions for further studies

An obvious limitation of the treatment is that the correction element in DI_2 is based on a simulation dataset based on real items related to an original dataset. It is not known how much the coefficients of 0.146 and 0.05 in the correction models depend on the original dataset. Controlled studies would be beneficial; they might suggest slightly different coefficients than 0.146 or 0.05. However, we note that the number of numerical sub-coefficients in the correction factor is minimal, and the estimates in the training dataset are based on 14,880 items with diverse characteristics and strong roots in the real-world test setting. Therefore, to some extent, the new coefficient appears to be relatively free from the original dataset.

Cross-validating the model by using data sets from the same base population and same test items did not challenge the models profoundly. This was done but the result was the same in both cases. Therefore, only the training data set was used in the empirical section. Control studies and replications of the design would be beneficial. It would be better to combine also the less studied estimators in the same simulation: RAC, EAC, D, G, D_2, G_2, R_{REG}, $DI20\%$, $DI27\%$, DI_2Log, and DI_2Lin.

Although the datasets used in the two simulations are relatively large and, from this point of view, convincing enough, the number of items with a wide scale in the datasets is small. In addition, the large-scale items in the training dataset were always associated with very short tests ($k=2$–4) and therefore the dataset is not strong if one wants to infer something specific about the behaviour of DI_2 from the number of items in the test. However, as the estimates are based on real test settings, they are likely to be applicable to real test settings. Systematic studies in this regard would obviously give us more knowledge about the behaviour of DI and DI_2, if not suggest a slightly different correction. In particular, such simulations where the item has more than seven degrees of freedom would enrich our knowledge of the coefficient.

Also, simulations regarding the possible over- and underestimation of the association in relation to the true discriminative power of the items in general would benefit us.

Appendix 1. Benchmarking estimators of item–score association

Product–moment correlation coefficient (Rit, PMC)

Rit, also known as point-biserial and point-polyserial correlation in the item analysis settings, is, technically, a product-moment correlation coefficient (PMC). In measurement modelling settings with an item g and a score X, *Rit* can be expressed as $\mathrm{PMC} = Rit = \rho_{gX} = \sigma_{gX}(\sigma_g \sigma_X)^{-1}$, where σ_{gX}refers to the item–score covariance and σ_g and σ_X to the standard deviations.

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

Polychoric correlation (RPC)

In the item analysis settings, *RPC* estimates the correlation between the unobserved latent, normal-assumed, continuous "item" and a "score" truncated to ordinal (or interval)-scaled forms. *RPC* is used to infer what *could be* the correlation between two normally distributed latent variables given the observed dataset. *RPC* differs from *Rit* from the viewpoint that we do not have a closed form for the estimation but there are several alternatives for the estimation process. One of these options is the two-step estimator by Martinson and Hamdan (1972), which is used in this article. Zaiontz (2025) presents a simplified form for this estimator: the task is to find the value of *Rit* that maximizes the log-likelihood function $LL = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij} LN\left(P\left(g = i, X = j\right)\right)$, where R and C refer to the number of categories in the item and score, respectively, n_{ij} refers to the number of cases in each cell of crosstable of g and X, and LN refers to natural logarithm taken during the process of each combination of i and j (see in-depth in Zaiontz, 2025).

In IBM SPSS software package, the syntax for *RPC* is not available although some macros are (see Lorenzo-Seva and Ferrando 2015). In SAS, the command PROC CORR provides *RPC*. Correspondingly, in the R environment, *RPC* can be calculated, for example, by the library "polychor" by Fox and Dusa (2022) and syntax Polychor(g, X) or by the library "DescTools" by Signorelli (2022) and there syntax *CorPolychor(g, X, ML=FALSE, control=list(), std.err=FALSE, maxcor=.9999)*## S3 method for class 'CorPolychor' *print(x, digits=max(3, getOption("digits") − 3)*,

...) (see https://rdrr.io/cran/DescTools/man/CorPolychor.html). For the empirical section, *RPC* was calculated manually by using Zaiont's (2025) syntax with a standard spreadsheet software (MS-Excel).

R-bireg and r-polyreg correlation (RREG)

By far, the best option for biserial (*RBS*) and polyserial (*RPS*) coefficients of correlation (Pearson 1909, 1913) seems to be a coefficient called r-biserial and r-polyreg correlation (*RREG*; originally developed by Lewis, Thayer, and Livingstone in a non-dated manuscript and discussed by Livinstone & Dorans, 2004 and Moses 2017). Assume that the observed item score (y_i) is determined by an underlying latent continuous variable θ manifested as the score variable X. The distribution of θ for test-takers with the observed value (x) in the score variable X is assumed to be normal with mean of βx and variance of 1, where β is an item parameter estimated by the probit regression model $P\left(y_i \leqslant 1 \,|x\right) = \Phi\left(a_i - \beta_i x\right)$ where Φ is the standard normal cumulative distribution function and a_i and β_i are intercept and slope parameters. After the estimate of β is computed ($\hat{\beta} = \beta$), *RREG* is calculated as $\rho_{REG} = \beta\sigma_X \Big/ \sqrt{\beta^2\sigma_X^2 + 1}$, where σ_X^2 is the population variance of the score variable X.

The estimate for β can be computed, 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.

For the empirical section, β and σ_X^2 were estimated by using SPSS software package and the estimates by *RREG* are computed with MS-Excel software.

Gamma (G) and dimension-corrected gamma (G_2)

Goodman–Kruskal gamma estimates the probability that two random pairs in two variables are in the same order (see formally in, e.g., Van der Ark and Van Aert 2015). The computational form of G uses the concepts of concordance and discordance related to the observed item scores of pairs of test-takers in $g\left(y_i, y_j\right)$ and $X\left(x_i, x_j\right)$ (see, e.g., Siegel and Castellan 1988). If y_i and y_j and corresponding x_i and x_j have ranks in the same direction, the pair is concordant and if they are in opposite order, the pair is discordant. Conventionally, the number of concordant pairs is denoted by P and the number of discordant pairs by Q. G proportions $P - Q$ with the number of pairs *where the direction is known* (compare later D): $G = (P - Q)/(P + Q)$.

As being an estimator based on probability with a linear nature, G tends to underestimate the item–score association in an obvious manner when the number of categories exceeds 4 (see the discussion and algebraic reasons in Metsämuuronen 2021a, 2022a). Hence, Metsämuuronen (2021a) suggests using dimension-corrected G (G_2) which brings the linear nature in G nearer the trigonometric nature of estimators based on covariance. The computational form of G_2 is $G_2 = G \times (1 + (1 - abs\,(G)) \times A)$where G is the observed value of G, *abs* refers to

absolute value, and $A = (1 - 1/df\,(g))^3$, where *df*(*g*) refers to the number of categories in the item minus 1. G_2 equals G with binary items and when the discrimination is deterministic ($G = G_2 = 1$) When the scale of item has more than two categories, the magnitude of the estimates by G_2 are always higher in comparison with those by G.

In IBM SPSS, the syntax CROSSTABS/TABLES = item BY Score/STATISTICS = GAMMA gives G. In SAS, the command PROC FREQ provides G by specifying the TEST statement by GAMMA, SMDCR options. Correspondingly, in the R environment G is calculated, for example, by the package "DescTool" by Signorelli (2022) with syntax *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 by MS-Excel software based on the observed G and *df*(*g*).

Delta (D) and dimension-corrected D (D_2)

Somers delta belongs to the same family as G and *Tau-b*. As does G, also D estimates the probability that two random pairs in two variables are in the same order. The computational form of D directed so that X explains the response pattern in g, the direction suitable for the item analysis settings (see discussion in Metsämuuronen 2020b, 2021a), is $D\,(g\,|X\,) = D = 2\,(P - Q)/\,(P + Q + 2T_g)$ where $2T_g$ refers to the number of tied pairs related to g.

Like G also D is an estimator based on probability with a linear nature and, hence, D tends to underestimate the item–score association in an obvious manner when the number of categories exceeds 3 (note 4 with G; see Metsämuuronen 2021a, 2022a). Hence, Metsämuuronen (originally 2020c, corrected in 2021a) proposes dimension-corrected D (D_2) which brings the linear nature in D nearer the trigonometric nature of estimators based on covariance. The computational form of D_2 has the same form as the one of G_2: $D_2 = D \times (1 + (1 - abs\,(D)) \times A)$where D is the observed value of $D\,(g\,|X\,)$and $A = (1 - 1/df\,(g))^3$as above with G.

In IBM SPSS, the syntax CROSSTABS/TABLES = item BY Score/STATISTICS = D gives D. In SAS, the command PROC FREQ provides D by specifying the TEST statement by D, SMDCR options. Correspondingly, in the R environment, D can be computed, for example, by package "DescTool" by Signorelli (2022) with the syntax *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*).

Attenuation-corrected Rit

Because *Rit* is known to be give estimates that are seriously attenuated or deflated when the scales of variables differ from each other as is the case in settings related to item analysis, Metsämuuronen (2022b, c) suggests a simple correction to *Rit* called the attenuation-corrected *R* (*RAC*). The attenuation-corrected *Rit* (*RAC*) is the ratio of the observed correlation ($\rho_{gX}^{Obs} = Rit$) and ρ_{gX}^{Max} given the observed data-

set: $RAC = \frac{\rho_{gX}^{Obs}}{\rho_{gX}^{Max}} = \frac{\sigma_{gX}^{Obs} \times (\sigma_g \sigma_X)^{-1}}{\sigma_{gX}^{Max} \times (\sigma_g \sigma_X)^{-1}} = \frac{\sigma_{gX}^{Obs}}{\sigma_{gX}^{Max}}$, where σ_{gX}^{Obs} and σ_{gX}^{Max} refer to the observed and maximal covariance between g and X.

There is no specific syntax for *RAC* available yet in the general software packages. However, the observed *Rit* can be computed by using the syntaxes given above. The maximal correlation is easy to compute by any general spreadsheet software package by sorting first the variables independently and then computing correlation between the sorted variables. In the R environment, *RAC* can be obtained by using the syntax cor(g, X)/cor(sort(g), sort(X)). For the empirical section, *RAC* was computed by using the MS-Excel software package.

Acknowledgements Sincere thanks to Counselor or Evaluation, Jukka Marjanen from Finnish Education Evaluation Center (FINEEC) for providing the syntax for *RREG*. The author sincere thanks the editorial office and two anonymous reviewers who suggested to do some additional analyses. These radically changed the tone in the text.

Funding statement Open Access funding provided by University of Turku (including Turku University Central Hospital). The author declares that no funds, grants, or other support were received during the preparation of this manuscript. The study has been completed as part of the EDUCA Flagship project funded by the Research Council of Finland (#358924, #358947).

Data availability The main dataset used in the article is available in in CSV format at https://doi.org/10.13140/RG.2.2.14808.16648 and in SPSS format at https://doi.org/10.13140/RG.2.2.24874.49602. The secondary dataset used in the article is available in CSV format at https://doi.org/10.13140/RG.2.2.14598.45127 and in SPSS format at https://doi.org/10.13140/RG.2.2.31375.66726.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

Ethic statement All necessary support and approvals are in place for the research.

Copyright statement All necessary copyrights are in place for the research.

Preprint Server https://doi.org/10.13140/RG.2.2.14430.20804.

References

Batanero CD (2007) *Suitability of teaching Bayesian inference in data analysis courses directed to psychologists.* [Unpublished doctoral dissertation]. University of Granada Spain. Available at https://iase-web.org/documents/dissertations/07.Diaz.pdf (Accessed Dec 27, 2022)

Bazaldua DAL, Lee Y-S, Keller B, Fellers L (2017) Assessing the performance of classical test theory item discrimination estimators in Monte Carlo simulations. Asia Pac Educ Rev 18(4):585–598. https://doi.org/10.1007/s12564-017-9507-4

Beuchert AK, Mendoza JL (1979) A monte carlo comparison of ten item discrimination indices. J Educ Meas 16(2):109–117

Bravais A (1844) Analyse Mathematique. Sur les probabilités des erreurs de situation d'un point. (Mathematical analysis. Of the probabilities of the point errors). Mémoires présentés Par Divers Savants à l'Académie Royale Des Siences De l'Institut De France (Memoirs Presented Var Scholars Royal Acad Sci Inst France) 9:255–332. https://books.google.fi/books?id=7g_hAQAACAAJ&redir_esc=y

Brennan RL (1972) A generalized upper-lower item discrimination index. Educ Psychol Meas 32(2):289–303. https://doi.org/10.1177/001316447203200206

Çelen Ü, Aybek EC (2022) A novel approach for calculating the item discrimination for likert type of scales. Int J Assess Tools Educ 9(3):772–786. https://doi.org/10.21449/ijate.1173356

Cook LL, Eignor DR, Taft HL (1988) A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. J Educ Meas 25(1):31–45

Cureton EE (1966a) Simplified formulas for item analysis. J Educ Meas 3(2):187–189. https://doi.org/10.1111/j.1745-3984.1966.tb00879.x

Cureton EE (1966b) Corrected item–test correlations. Psychometrika 31(1):93–96. https://doi.org/10.1007/BF02289461

D'Agostino RB, Cureton EE (1975) The 27% rule revisited. Educ Psychol Meas 19(1):47–50. https://doi.org/10.1177/001316447503500105

Ebel RL (1954a) How an examination service helps college teachers to give better tests. *Proceedings of the 1953 invitational conference on testing problems.* Educational Testing Service

Ebel RL (1954b) Procedures for the analysis of classroom tests. Educ Psychol Meas 14(2):352–353. https://doi.org/10.1177/001316445401400215

Ebel RL (1965) Measuring educational achievement. Prentice-Hall

Ebel RL (1967) The relation of item discrimination to test reliability. J Educ Meas 4(3):125–128. https://doi.org/10.1111/j.1745-3984.1967.tb00579.x

Ebel RL (1972) *Essentials* of *Educational Measurement (*1st Edition). Prentice Hall

Ebel RL (1979) *Essentials of educational measurement* (3rd Edition). Prentice-Hall

Ebel RL, Frisbie DA (1986) *Essentials of Educational Measurement* (5th Edition). Pearson

Englehart MD (1965) A comparison of several item discrimination indices. Educ Psychol Meas 2(1):69–76. https://doi.org/10.1111/j.1745-3984.1965.tb00393.x

ETS (1960) Short-cut statistics for teacher-made tests. Educational Testing Service

ETS (2022) *Glossary of standardized testing terms*. Educational Testing Service. Available at https://www.marylandpublicschools.org/Documents/commissiononassessments/ETSGlossaryStandardizedTestingTerms.pdf (Accessed Dec 27, 2022)

Fan X (1998) Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educ Psychol Meas 58(3):357–381. https://doi.org/10.1177/00131644980580030

Feldt LS (1963) Note on use of extreme criterion groups in item discrimination analysis. Psychometrika 28(1):97–104. https://doi.org/10.1007/BF02289553

Fennema E, Sherman JA (1976) Fennema-Sherman mathematics attitude scales: instruments designed to measure attitudes toward the learning of mathematics by females and males. J Res Math Educ 7(5):324–326. https://doi.org/10.2307/748467

FINEEC (2018) Unpublished dataset opened for the re-analysis 18.2.2018. National assessment of learning outcomes in mathematics at grade 9 in 2004. Finnish National Education Evaluation Centre

Forlano G, Pinter R (1941) Selection of upper and lower groups for item validation. J Educ Psychol 32(7):544–549. https://doi.org/10.1037/h0058501

Fox J, Dusa A (2022) Package polycor. Polychoric and Polyserial Correlations. https://cran.r-project.org/web/packages/polycor/polycor.pdf

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

Goodman LA, Kruskal WH (1954) Measures of association for cross classifications. J Am Stat Assoc 49(268):732–764. https://doi.org/10.2307/2281536

Goodman LS, Kruskal WH (1959) Measures of association for cross classification. Part II: further discussion and references. J Am Stat Assoc 54:123–163. https://doi.org/10.2307/2282143

Hales LW (1972) Method of obtaining the index of discrimination for item selection and selected test characteristics: a comparative study. Educ Psychol Meas 32(4):929–937. https://doi.org/10.1177/001316447203200407

Hambleton RK, Jones RW (1993) Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues Pract 12(3):38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Harris CW, Wilcox RR (1980) Brennan's b is Peirce's theta. Educ Psychol Meas 40(2):307–311. https://doi.org/10.1177/001316448004000204

Henrysson S (1963) Correction of item–total correlations in item analysis. Psychometrika 28(2):211–218. https://doi.org/10.1007/BF02289618

Henrysson S (1971) Gathering, analyzing and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd edition) (pp. 130–159). American Council on Education

IBM (2017) *IBM SPSS Statistics 25 Algorithms*. IBM. https://www.ibm.com/docs/en/SSLVMB_25.0.0/pdf/en/IBM_SPSS_Statistics_Algorithms.pdf (Accessed Dec 27, 2022)

Johnson AP (1951) Notes on a suggested index of item validity: the U-L index. J Educ Psychol 42(8):499–504. https://doi.org/10.1037/h0060855

Kelley TL (1939) The selection of upper and lower groups for the validation of test items. J Educ Psychol 30(1):17–24. https://doi.org/10.1037/h0057123

Kelley T, Ebel R, Linacre JM (2002) Item discrimination indices. Rasch Meas Trans 16(3):883–884

Kendall MG (1945) The treatment of ties in ranking problems. Biometrika 33(3):239–251. https://doi.org/10.2307/2332303

Kendall MG (1948) *Rank correlation methods* (1st Edition). Charles Griffin & Co Ltd

Kohli N, Koran J, Henn L (2015) Relationships among classical test theory and item response theory frameworks via factor analytic models. Educ Psychol Meas 75(3):389–405. https://doi.org/10.1177/0013164414559071

Lawson S (1991) One parameter latent trait measurement: do the results justify the effort? In: Thompson B (ed) Advances in educational research: substantive findings, methodological development, vol 1. JAI, pp 159–168

Lewis C, Thayer D, Livingstone SA (n.d.). *A regression-based polyserial correlation coefficient.* Unpublished manuscript

Liu F (2008) Comparison of several popular discrimination indices based on different criteria and their application in item analysis. Univ Ga. https://getd.libs.uga.edu/pdfs/liu_fu_200808_ma.pdf

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

Long JA, Sandiford P (1935) *The validation of test items (Bulletin No. 3).* Department of Educational Research University of Toronto

Lord FM, Novick MR, Birnbaum A (1968) Statistical theories of mental test scores. Addison–Wesley Publishing Company

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

Macdonald P, Paunonen SV (2002) A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. Educ Psychol Meas 62(6):921–943. https://doi.org/10.1177/0013164402238082

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

Martinková P, Drabinová A (2018) Shinyitemanalysis for teaching psychometrics and to enforce routine analysis of educational tests. R J 10(2):503–515

Martinková P, Štěpánek L, Drabinová A, Houdek J, Vejražka M, Štuka Č (2017) Semi-real-time analyses of item characteristics for medical school admission tests. In M. Ganzha, L. Maciaszek, M. Paprzycki (eds), *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, ACSIS, Vol. 11*, 189–194. https://doi.org/10.15439/2017F380

Martinson EO, Hamdan MA (1972) Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. J Stat Comput Simul 1(1):45–54. https://doi.org/10.1080/00949657208810003

McGrath RE, Meyer GJ (2006) When effect sizes disagree: the case of *r* and *d*. Psychol Methods 11(4):386–401. https://doi.org/10.1037/1082-989x.11.4.386

Mehrens WA, Lehmann IJ (1973) *Measurement and evaluation in education and psychology* (1st Edition). Holt, Rinehart and Winston

Mehrens WA, Lehmann IJ (1991) *Measurement and evaluation in education and psychology* (4th Edition). Harcourt Brace College Publishers

Metsämuuronen J (2012) Challenges of the Fennema-Sherman test in the international comparisons. Int J Psychol Stud 4(3):1–22

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.worldwidejournals.com/global-journal-for-research-analysis-GJRA/fileview/November_2016_1478701072__159.pdf

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

Metsämuuronen J (2020a) Generalized discrimination index. Int J Educational Methodol 6(2):237–257. https://doi.org/10.12973/ijem.6.2.237

Metsämuuronen J (2020b) Somers *D* as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. Int J Educational Methodol 6(1):207–221. https://doi.org/10.12973/ijem.6.1.207

Metsämuuronen J (2020c) Dimension-corrected somers d for the item analysis settings. Int J Educational Methodol 6(2):297–317. https://doi.org/10.12973/ijem.6.2.297

Metsämuuronen J (2021a) Goodman–kruskal gamma and dimension-corrected gamma in educational measurement settings. Int J Educational 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) Essentials of visual diagnosis of test items. Logical, illogical, and anomalous patterns in tests items to be detected. *Practical Assessment, Research, and Evaluation, PARE*, 27, Art. 5. https://doi.org/10.7275/n0kf-ah40

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

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

Metsämuuronen J (2022d) Directional nature of the product–moment correlation coefficient and some consequences. Front Psychol 13:988660. https://doi.org/10.3389/fpsyg.2022.988660

Metsämuuronen J (2022e) Artificial systematic Attenuation in Eta squared and some related consequences. Attenuation-corrected Eta and Eta squared, negative values of Eta, and their relation to pearson correlation. Behaviormetrika. https://link.springer.com/content/pdf/ https://doi.org/10.1007/s41237-022-00162-2.pdf

Metsämuuronen J (2022f) Attenuation-corrected reliability and some other MEC-corrected estimators of reliability. Appl Psychol Meas. https://doi.org/10.1177/01466216221108131

Metsämuuronen J (2022g) Item-rest correlations revisited. Algebraic reasons why the estimates by item-rest correlation are more deflated than those by item-test correlation, and some coefficients to consider as alternatives. Preprint available at https://doi.org/10.13140/RG.2.2.24704.71687 (Accessed Dec 27, 2022)

Metsämuuronen J (2022h) Deflation-corrected estimators of reliability. Front Psychol 12:748672. https://doi.org/10.3389/fpsyg.2021.748672

Metsämuuronen J (2022i) Typology of deflation-corrected estimators of reliability. Front Psychol 13:891959. https://doi.org/10.3389/fpsyg.2022.891959

Metsämuuronen J (2023) *Matematiikkaa COVID-19-pandemian varjossa III. Syventäviä analyyseja matematiikan 9. luokan arvioinnista keväällä 2021*. [Mathematics in the Shadow of the COVID-19 Pandemic III. Deepening analysis of the achievement in mathematics at the end of 9th grade in Spring 2021] Publications 31:2023. Finnish Education Evaluation Centre, FINEEC. https://www.karvi.fi/sites/default/files/sites/default/files/documents/KARVI_3123.pdf [In Finnish, English Abstract]

Metsämuuronen J, Nousiainen S (2021) *Matematiikkaa COVID-19-pandemian varjossa. Matematiikan osaaminen 9. luokan lopussa keväällä 2021*. [Mathematics in the Shadow of the COVID-19 Pandemic. Achievement in mathematics at the end of 9th grade in Spring 2021] Publications 27:2021. Finnish Education Evaluation Centre, FINEEC. https://www.karvi.fi/sites/default/files/sites/default/files/documents/KARVI_2721.pdf [In Finnish, English Abstract]

Metsämuuronen J, Nousiainen S (eds) (2023) *Matematiikkaa COVID-19-pandemian varjossa II. Menetelmälliset ratkaisut matematiikan 9. luokan arvioinnissa keväällä 2021.* [Mathematics in the Shadow of the COVID-19 Pandemic II. Methodological solutions of the assessment of mathematics at the 9th grade in Spring 2021] Publications 5:2023. Finnish Education Evaluation Centre, FINEEC. https://www.karvi.fi/sites/default/files/sites/default/files/documents/KARVI_0523.pdf [In Finnish, English Abstract]

Metsämuuronen J, Tuohilampi L (2014) Changes in achievement in and attitude toward mathematics of the Finnish children from grade 0 to 9—a longitudinal study. J Educ Dev Psychol 4(2):145–169

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

Ndalichako JL, Rogers WT (1997) Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. Educ Psychol Meas 57(4):580–589. https://doi.org/10.1177/0013164497057004005

Nitko AJ, Hsu T, annual meeting of the American Educational Research Association (1984) *Item analysis appropriate for domain-referenced classroom testing*. Paper presented at the (68th, New Orleans, LA, April 23–27, 1984). Available at https://files.eric.ed.gov/fulltext/ED242781.pdf (Accessed Dec 27, 2022)

Okada K (2013) Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. Behaviormetrika 40:129–147. https://doi.org/10.2333/bhmk.40.129

Okada K (2017) Negative estimate of variance-accounted-for effect size: how often it is obtained, and what happens if it is treated as zero. Behav Res Methods 49:979–987. https://doi.org/10.3758/s13428-016-0760-y

Olsson U (1980) Measuring correlation in ordered two-way contingency tables. J Mark Res 17(3):391–394. https://doi.org/10.1177/002224378001700315

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 part III. Regression, heredity, and panmixia. Philosophical Trans Royal Soc Lond Ser Containing Papers Math or 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. Philosophical Trans Royal Soc 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. Philosophical Trans Royal Soc 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.

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

Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche

Rasch G Review of the cooperation of Professor B. D. Wright, University of Chicago, and, Professor G, Rasch (1972) University of Copenhagen; letter of June 18, 1972. *Rasch Measurement Transactions*, 1988, *2*(2), 19. http://www.rasch.org/rmt/rmt22c.htm

Rizzo M, Székely G (2022) Package Energy. E-Statistics: Multivariate inference via the energy of data. https://cran.r-project.org/web/packages/energy/index.html (Accessed Dec 27, 2022)

Ross J, Lumsden J (1964) Comment on feldt's use of extreme groups. Psychometrika 29(2):207–209. https://doi.org/10.1007/BF02289701

Ross J, Weitzman RA (1964) The twenty-seven per cent rule. Ann Math Stat 35(1):214–221. https://doi.org/10.1214/aoms/1177703745

Ruscio J (2008) A probability-based measure of effect size: robustness to base rates and other factors. Psychol Methods 13(1):19–30. https://doi.org/10.1037/1082-989X.13.1.19

Shannon GA, Cliver BA (1987) An application of item response theory in the comparison of four conventional item discrimination indices for criterion-referenced tests. J Educ Meas 24(4):347–356. https://doi.org/10.1111/j.1745-3984.1987.tb00285.x

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

Signorelli A (2022) Package DescTools. Tools for Descriptive Statistics and Exploratory Data Analysis. https://rdrr.io/cran/DescTools/

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

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101. https://doi.org/10.2307/1422689

Stata corp (2018) *Stata manual*. Stata. https://www.stata.com/manuals13/mvalpha.pdf

Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35(6):2769–2794. https://doi.org/10.1214/009053607000000505

Tristan LA (1998) The item discrimination index: does it work? Rasch Meas Trans 12(1):626

Van der Ark LA, Van Aert RCM (2015) Comparing confidence intervals for Goodman and Kruskal's gamma coefficient. J Stat Comput Simul 85(12):2491–2505. https://doi.org/10.1080/00949655.2014.932791

Whitney DR, Sabers DL (1971) Two generalizations of the item discrimination index to multi-score items. J Exp Educ 39(3):88–92

Wiersma W, Jurs SG (1985) *Educational measurement and testing* (1st Edition). Allyn and Bacon

Wiersma W, Jurs SG (1990) *Educational measurement and testing* (2nd Edition). Allyn and Bacon

Wolf R (1967) Evaluation of several formulae for correction of item-total correlations in item analysis. J Educ Meas 4(1):21–26. https://doi.org/10.1111/j.1745-3984.1967.tb00565.x

Yi-Hsin C, Li I (2015) IA_CTT: A SAS® macro for conducting item analysis based on classical test theory. Paper CC184. https://analytics.ncsu.edu/sesug/2015/CC-184.pdf

Zaiontz C (2022) *Real Statics Using Excel.* Polychoric Correlation using Solver. http://www.real-statistics.com/correlation/polychoric-correlation/polychoric-correlation-using-solver/ (Accessed Dec 27, 2022)

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