Research Article doi: 10.12973/ijem.6.2.297

International Journal of Educational Methodology
Volume 6, Issue 2, 297 - 317.
ISSN: 2469-9632
http://www.ijem.com/

Dimension-Corrected Somers' D for the Item Analysis Settings

Jari Metsämuuronen*
Finnish Education Evaluation Centre,
FINLAND

NLA University College,
NORWAY

Received: February 17, 2020 ▪ Revised: April 5 2020 ▪ Accepted: April 21, 2020

Abstract: A new index of item discrimination power (IDP), dimension-corrected Somers' D (D2) is proposed. Somers' D is one of the superior alternatives for item–total- (Rit) and item–rest correlation (Rir) in reflecting the real IDP with items with scales 0/1 and 0/1/2, that is, up to three categories. D also reaches the extreme value +1 and −1 correctly while Rit and Rir cannot reach the ultimate values in the real-life testing settings. However, when the item has four categories or more, Somers' D underestimates IDP more than Pearson correlation. A simple correction to Somers' D in the polytomous case seems to lead to be effective in item analysis settings. In the simulation with real-life items, D2 showed very few cases of obvious underestimation and practically no cases of obvious overestimation. With certain restrictions discussed in the article, D2 seems to be a good alternative for these classic estimators not only with dichotomous items but also with the polytomous ones. In general, the magnitudes of the estimates by D2 are higher than those by Rit, Rir, and polychoric correlation and they seem to be close of those of bi- and polyserial correlation coefficients without out-of-range values.

Keywords: *Item analysis, Pearson correlation, item–total correlation, item–rest correlation, Somers' D, item discrimination power.*

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

Introduction

Item discrimination and the deterministic pattern

Item discrimination power (IDP)—one of the three essential parameters of a test item—is classically defined as the efficiency of a single item to discriminate between lower- and higher-scoring test-takers (see Educational Testing Service [ETS], 2020; Liu, 2008; Lord & Novick, 1968; MacDonald & Paunonen, 2002). Metsämuuronen (2020a) notes that this loose definition is not very practical while assessing the possible under- and overestimation produced by different estimators of IDP in the real-life settings. Hence, he discusses an operational definition of IDP related to the concept of *deterministic* item discrimination. Deterministic item discrimination refers to the pattern in which the score explains perfectly the behavior in the item, and then we expect to see the perfect explaining power between two variables ($\rho^2_{XY}=1$) that implies the perfect association ($\rho_{XY}=1$). In other words, when the latent trait can predict the behavior of the test-takers in the test item in a deterministic manner the test item is ultimately reliable. In practical settings related to item analysis, the perfect explaining power is achieved when the *order* of the cases both in the item and the score are identical. Hence, Metsämuuronen defines the ultimate IDP as a condition where "after arranging the test-takers by the score, or the measurement scale, the item can discriminate the lower performing test-takers from the higher performing test-takers in a deterministic manner" (Metsämuuronen, 2020a, p. 208).

When it comes to detecting the ultimate IDP, two widely used classical estimators of IDP, item–total correlation (ρ_{gX} , *Rit*; based on Pearson, 1896) and item–rest correlation (ρ_{gP} , *Rir*; Henrysson, 1963), are not strong; they cannot reach the ultimate value $\rho_{XY}=1$ because of the mismatch of the dimensions of the item and the score, and the underestimation of IDP may be drastic if the item difficulty is extreme (e.g. Metsämuuronen, 2016; 2017a; 2020a;

* **Correspondence:**
Jari Metsämuuronen, Finnish Education Evaluation Centre, P.O. Box 28, FI-00101 Helsinki, Finland. ✉ jari.metsamuuronen@gmail.com

2020b). Two other estimators of IDP, bi- and polyserial correlation coefficients, by using standard procedures of estimation (see Drasgow, 1986), tend to give obvious overestimates (out-of-range values) when ρ_{gX} and item variance are high (e.g. Lord & Novick, 1968). This is specifically true with the deterministic patterns with non-normal or even distribution in the score (see more recent literature and examples in Metsämuuronen, 2020a; 2020b). Also one of the superior alternatives to ρ_{gX} and ρ_{gP}, polychoric correlation (Pearson, 1900; 1913), cannot reach the ultimate IDP by using standard procedures of estimation because of the technical reasons. In the deterministic patterns, one of the "superior alternatives" to ρ_{gX}, Somers' *D* (Somers, 1962; see Metsämuuronen, 2020a), reaches correctly the values +1 and −1, it is stable with extreme values (Newson, 2002), and it gives estimates for IDP that are remarkably closer the real IDP than *Rit* and *Rir* (Metsämuuronen, 2020a). These are advances in the practical educational testing settings, when the sample sizes may be small and the normality in the score cannot be ensured (see examples in Aslan & Aybek, 2019; Delil & Ozcan, 2019). Hence, the characteristics of Somers' *D* are worth of studying in measurement modeling settings.

Statistical model latent to Somers' D

Assume two ordinal observed variables *g* (item) and *X* (score) that have *r* and *s* distinctive categories, respectively. Within the measurement modelling settings, the observed values g_i and x_j are driven by a continuous latent variable θ common to both variables. The threshold values of θ for each category in *g* are denoted by γ_i and in *X* by τ_j. Then, the variable *g* is related to θ so that $g = g_i$, if $\gamma_{i-1} \leq \theta < \gamma_i$, i = 1, 2,..., R and $X = x_j$, if $\tau_{j-1} \leq \theta < \tau_j$, j = 1, 2, ..., S as illustrated in Figure 1. We define that $\gamma_0 = \tau_0 = -\infty$ and $\gamma_R = \tau_S = +\infty$, and we assume that $g_1 < g_i < g_r$ and $x_1 < x_j < x_s$.

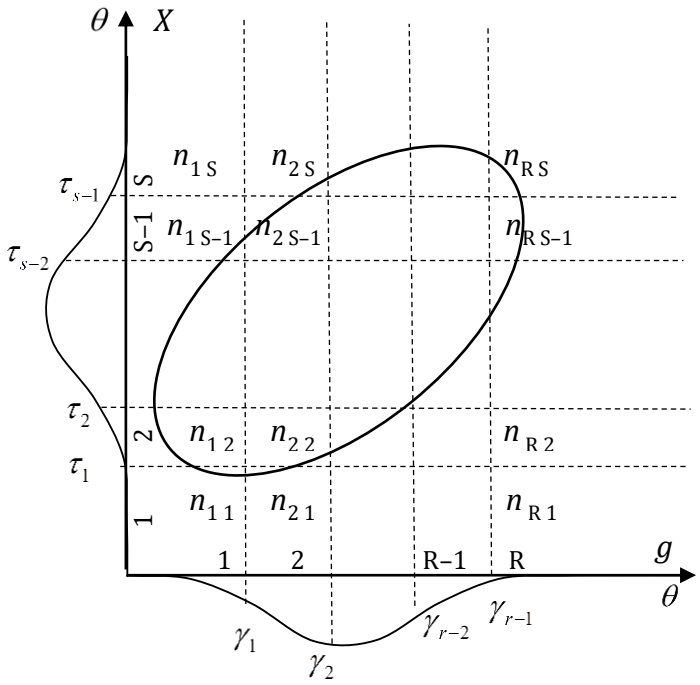

Figure 1. A latent variable θ categorized into two ordinal scales and the number of times the observation (g_i, x_j) is obtained in the sample (n_{gX})

From the traditional viewpoint of correlation coefficients related to the item analysis, the observed correlation between the interval-scaled variables *g* and *X* is item–total correlation (ρ_{gX}), the observed correlation between the binary *g* and ordinal *X* is rank-biserial correlation (ρ_{RB}), the inferred correlation between the latent θ and observed *X* is polyserial correlation ($\rho_{\theta X}$), and the inferred correlation between two latent variables is polychoric correlation ($\rho_{\theta\theta}$). In the last, within the measurement modelling settings, we would expect to obtain perfect correlation, and the further $\rho_{\theta\theta}$ is from 1, the more measurement error is included in the measurement instruments, including both items and the score. We note that the correlational viewpoint to the item discrimination is based on *covariance* between the item and the score.

From the viewpoint relevant with this article, the family of Somers' *D*, including Kendall's tau-a and tau-b (Kendall, 1938), and Goodman–Kruskal *G* (Goodman & Kruskal, 1954), item discrimination is approached from the *probability*

viewpoint. Somers' *D* estimates the probability (π) that two randomly chosen pair of test-takers have the same order in both the item and score (see Van der Ark & Van Aert, 2015). The probability for the same order is

$$\pi_P = \sum_{r=1}^{R}\sum_{c=1}^{C}\pi_{rc}\left(\sum_{i>r}\sum_{j>c}\pi_{ij} + \sum_{i<r}\sum_{j<c}\pi_{ij}\right) \quad (1a)$$

and the probability for the opposite order is

$$\pi_Q = \sum_{r=1}^{R}\sum_{c=1}^{C}\pi_{rc}\left(\sum_{i>r}\sum_{j<c}\pi_{ij} + \sum_{i<r}\sum_{j>c}\pi_{ij}\right) \quad (1b)$$

The probabilities of tied pairs related to rows and columns are π_{T_R} and π_{T_C}, respectively. The latent δ proportions the probabilities of *P* and *Q* with *maximal possible* number of pairs to the *same* direction (including also the tied pairs). Hence, the relevant direction related to the article, that is, the latent δ conditioned so that the column factor explains the row factor is defined as

$$\delta = \frac{\pi_P - \pi_Q}{\pi_P + \pi_Q + \pi_{T_R}}. \quad (2)$$

Somers' D in the practical item analysis settings

Somers' *D* approximates the latent *δ*. The computational forms of Somers' *D* are usually expressed by using the concepts of concordance and discordance between the values of *g* and *X*. By using the concepts of *P* and *Q*, the specific coefficient relevant to item analysis, *D* given *g* in condition of *X*, that is, Somers' $D(g|X)$†, has a simplified form of

$$D(g|X) = D = \frac{2(P-Q)}{N^2 - \sum_{i=1}^{R}\left(n_i^{\,2}\right)} \quad (3)$$

where n_{gi} is the number of cases in the categories $g = i$ related to item *g* and $P = \sum_{i,j} n_{ij}N_{ij}^{+}$ and $Q = \sum_{i,j} n_{ij}N_{ij}^{-}$ (Metsämuuronen, 2017b; Siegel & Castellan, 1988; note that, in the literature related to Somers' *D*, this is notated as $D(X|g)$). N_{ij}^{+} refers to the number of pairs in the cells below and to the *right* of the cell n_{ij}. Correspondingly, N_{ij}^{-} refers to the number of pairs in the cells below and to the *left* of the cell n_{ij}. The form is simplified because the values of *P* and *Q* are calculated only in one direction; more complicated form related to Eqs. (1) and (2) is seen in Section "Asymptotic sampling variance and standard error". The statistical properties of Somers' *D* have been discussed, for example, by Agresti (2010), Newson (2002; 2006; 2008) and Siegel and Castellan (1988), and practical procedures, for example, by Metsämuuronen (2017b).

Because of Eq. (3), Somers' $D(g|X)$ tells the proportion of the logically ordered test-takers in the item after the cases are ordered by the score. This fits well with the definition by Metsämuuronen (2020a) related to IDP. As does the correlation coefficient, $D(g|X)$ varies between –1 and +1. In the item analysis settings, the value $D(g|X)$ = +1 indicates the positive deterministic pattern: after ordered by the score, all the test-takers in the higher-ranked subsample(s) *j* in the item are (correctly) ranked higher than those in the lower subsample(s) *i*. The value $D(g|X)$ = –1 indicates the ultimately pathological situation that all the cases in the *lower* subsample(s) *i* would be ranked *higher* than those in the higher subsample *j*. The value $D(g|X)$ = 0 refers to a situation that the number of correctly ordered

† It is good to note the seemingly confusing notation related to Somers' *D* pointed by Metsämuuronen (2020a). In the traditional settings of conditions, the direction of condition $(g|X)$ usually means "*g* in condition of *X*", that is, "*g* is dependent on *X*", that is *"g* dependent". However, within the notation related to Somers' *D*, $D(X|g)$ is called "*g* dependent" (see Metsämuuronen, 2017b; Newson, 2002; 2006; 2008; Siegel & Castellan, 1988). In this article, the specific notation $D(g|X)$ refers to "*g* dependent" which, in the outputs of some generally known software packages such as IBM SPSS as well as *R* libraries, would be called "score dependent". See the practical notes of this notation in relation to the estimates in Metsämuuronen (2020a).

("concordant") test-takers equals the incorrectly ordered ("discordant") test-takers and, hence, the item cannot discriminate the test-takers from each other at all. Basically, the interpretation in the magnitude of the estimates by $D(g|X)$ is the same as that in ρ_{gX} with the note that, in real-life datasets, ρ_{gX} cannot reach perfect +1 or –1 while $D(g|X)$ can.

By using a comparison with real-life items, Metsämuuronen (2020a) showed that Somers' $D(g|X)$, (*D* henceforward), would be a good alternative for the generally used classical estimators of IDP. This is specifically true with binary items in relation with ρ_{gX} and ρ_{gP} as well as the family of bi- and polyserial correlations (ρ_{BS}, ρ_{PS}) and the polychoric correlation coefficient (ρ_{PC}) (Pearson, 1900; 1913). In comparison with ρ_{gX} and ρ_{gP}, *D* underestimates IDP less and is stable also with the items with extreme difficulty, which ρ_{gX} and ρ_{gP} may radically underestimate the IDP of. In comparison with ρ_{BS} and ρ_{PS}, *D* does not give obvious overestimates nor obvious underestimates as ρ_{BS} and ρ_{PS} may easily give. In comparison with ρ_{PC}, *D* relates with the *known* composite of items and score, and this information can be used in further analysis while ρ_{PC} refers to an unknown, unreachable, and hypothetical composites that are difficult to use in the analysis. In comparison with some other directional coefficients such as Goodman–Kruskal lambda and tau (Goodman & Kruskal, 1954) or Pearson's eta coefficient (η) (Pearson, 1903, 1905), *D* can detect the ultimate discrimination in the *item* while lambda, tau, and eta can detect the ultimate discrimination in the *score.* (Metsämuuronen, 2020a.)

Although *D* seems to be a "superior alternative" for ρ_{gX} and ρ_{gP} in the binary case, in the comparison by Metsämuuronen (2020a), *D* appeared to face a major practical challenge relevant to polytomous items. Although *D* reaches the ultimate values of IDP accurately, the estimates underestimate the IDP in an obvious manner when the number of categories in the marginal distribution of the item exceeds three and when the discrimination is not perfect or near perfect (Metsämuuronen, 2020a; see also Goktas & Isci; 2011; Newson, 2002). This is elaborated in what follows.

Underestimation in D in the empirical datasets

Metsämuuronen (2020a) noted the obvious patterns of underestimation in *D* with real-world datasets. The underestimation is strictly related to the number of categories in the items scale, that is, to the degrees of freedom of the marginal distribution of the item ($df(g) = r - 1$). When the number of marginal categories in the item exceeds three ($df(g) > 2$), ρ_{gX} appears to be superior to *D* reflecting IDP (Figure 2).

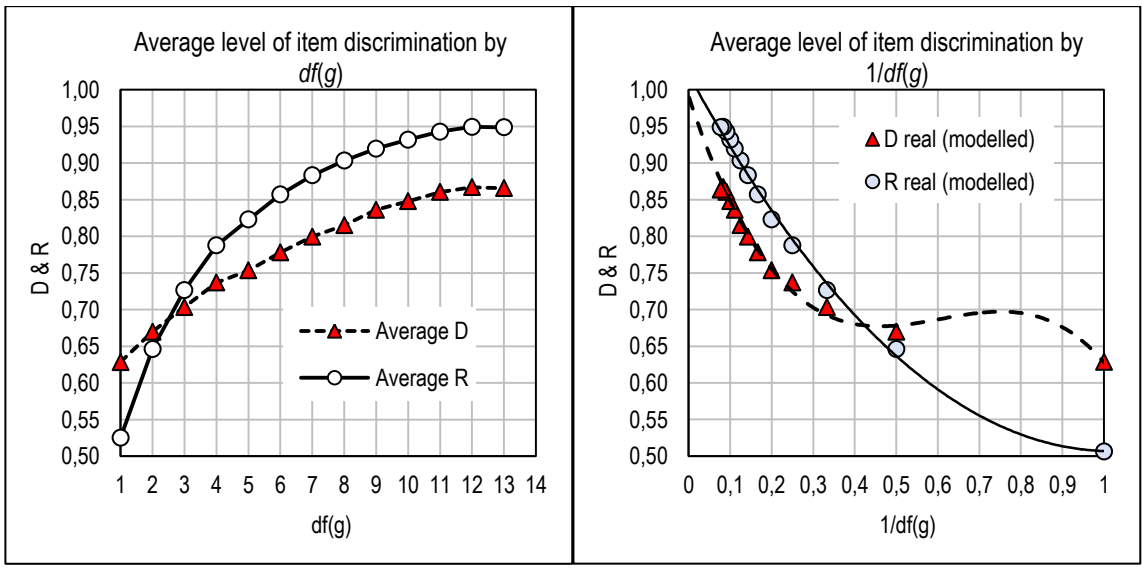

Figure 2. Underestimation in D in relation with ρ_{gX} (R) as a function of df(g) and 1/df(g)

The right-hand side graph in Figure 2 illustrates a practical peculiarity embedded in ρ_{gX} as well as in *all* estimators of IDP in item analysis settings, that the estimate approximates perfect 1 the less there are items in the test and the more there are categories in the items. The phenomenon is obvious when we recall that, in the measurement modeling

settings, the latent variable θ is common for both the item and the score (see Figure 1), and that the association of item g and score X is determined mechanically because the score is a compound of the items. The latter was the reason why Henrysson (1963) suggested his procedure (*Rir*); ρ_{gX} is characterized as "spuriously" inflated (e.g., Cureton, 1966, p. 93; Howard & Forehand, 1962, p. 731; Wolf, 1967, p. 21). When we think about a "test" with only *one* item: the correlation between the item and the "score" formed by this item, would be, obviously, perfect $\rho_{gX} = D = 1$. Correspondingly, the more we have items comprising the test score the further Pearson correlation between a single item and the score tends to deviate from 1 even if the score would explain perfectly the behavior in the item. Obviously, this phenomenon of approaching the value 1 does not make sense outside the measurement modeling settings but, in what follows, this plays a significant role in deriving the dimension correction to Somers' *D*.

Underestimation in Somers' D with polytomous items from the theoretical viewpoint

Although *D* underestimates IDP in obvious manner, the interpretation of the matter is somewhat challenging because PMC and *D* tell about different information of the relation of the item and the score discussed above. While ρ indicates *covariation* between the item and score, *D* indicates *probability* that two randomly chosen test-takers have the same order in both the item and score or the *proportion* of logically ordered test-takers in the item after they are ordered by the score. Anyhow, underestimation in *D* in relation to ρ_{gX} is expected because of Greiner's relation (Greiner, 1909) related to the connection of Kendall Tau-a, Somers' *D,* and Pearson correlation discussed by Kendall (1949) and Newson (2002). Assuming two independent variables *X* and *Y* with continuous scales (implying no ties) sampled from a bivariate normal distribution, Kendall Tau-a equals Somers' *D*. Then, Greiner's relation gives the association between *D* and ρ_{XY} as:

$$\rho_{XY} = \sin\left(\frac{\pi}{2} D\right). \tag{4}$$

From Eq. (4) we know that the values by *D* of 0, $\pm 1/3$, $\pm 1/2$, and ± 1 correspond to the values by ρ_{XY} of 0, $\pm 1/2$, $\pm 1/\sqrt{2}$, and ± 1, respectively (see Figure 3). Then, in the case of two normally distributed continuous variables, except for the extreme values ± 1 and 0, the magnitude of ρ_{XY} is greater than that of *D*. Consequently, because of Eq. (4), and because ρ_{gX} always underestimates association, the estimates by *D* are expected to underestimate IDP more than ρ_{gX} when the estimate by *D* differs from 0 and ± 1 and the number of marginal categories in the item is high.

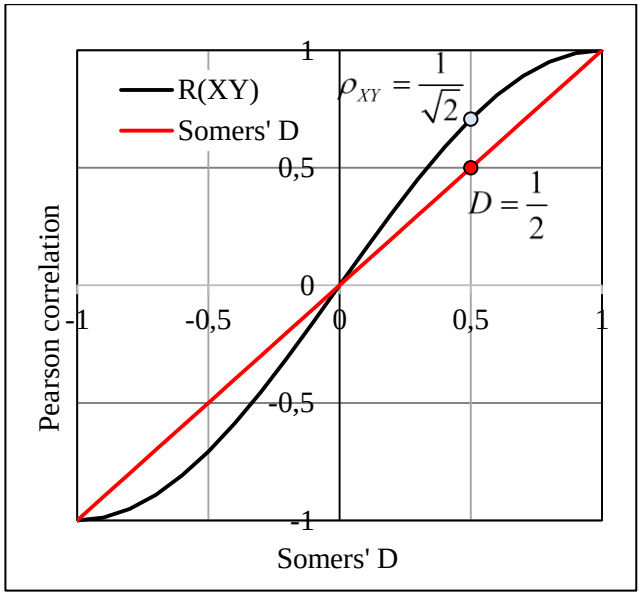

Figure 3. Relation of Pearson correlation (R_{XY}) and Somers' D with continuous variables X and Y

Because of the obvious disadvantage in *D* with the polytomous items to underestimate IDP even more than ρ_{gX}, Metsämuuronen (2020a) suggests that a "dimension-corrected Somers' *D*" could be worth of deriving. While *D* is a "superior alternative" to ρ_{gX} and ρ_{gP} in binary datasets, "dimension-corrected *D*" could be a "superior alternative" in the polytomous cases. As far it is known, such correction has not been proposed yet. The aim of the article is to derive a dimension-corrected version of *D* for the measurement modelling settings to reduce the obvious underestimation

obtained by *D* in the polytomous datasets.

Research questions

This article derives a dimension-correction version of *D* for the item analysis settings. After the derivation, the following questions are asked:

1) What are the general characteristics of the new coefficient in comparison with ρ_{gX}, ρ_{gP}, ρ_{BS}, ρ_{PS}, and ρ_{PC}?
2) What is the sampling variance and standard error of the new coefficient?
3) To what extent the new coefficient produces obvious underestimates?
4) To what extent the new coefficient produces obvious overestimates?

Methodology

Research design

The course of the study starts by deriving a "dimension-corrected *D*". This is done by modelling the error in *D* in 1,296 datasets with different number of test-takers (*N*), test lengths (*k*), difficulty levels ($\bar{p}$), reliabilities (α), and degrees of freedom in the item $df(g) = r - 1$, and in the score $df(X) = s - 1$. The datasets and items are presented in the next section.

After the derivation of the new coefficient, the asymptotic sampling variance and standard error are derived and a numerical example of the use of the coefficient is given with the comparison with the relevant benchmark coefficients.

The general characteristics of the new coefficient including the behavior in the extreme datasets, its limits as well as the potential over- and underestimation estimation are studied.

Finally, the advantages, limitations and possible ways to utilize the coefficient are discussed and suggestions for the further studies are given.

Datasets used in the derivation

The dimension correction to *D* is derived by using 13,392 real-world items from 1,296 datasets and the knowledge of the pattern of underestimation related to $df(g)$ illustrated in Figure 2. The datasets used in the derivation are formed by different combinations of randomly selected test-takers from a national-level dataset of 4,000 test-takers of a mathematics test for grade 9 with 30 binary items (Finnish Education Evaluation Centre [FINEEC], 2018). The difficulty levels of the items in the original dataset ranged from $0.24 < p < 0.95$ with the average difficulty level of $\bar{p} = 0.63$, the item discrimination ranged from $0.332 < \rho_{gX} < 0.627$ with the average $\bar{\rho}_{gX} = 0.481$, and the lower bound for the reliability was $\alpha = 0.885$ and, if using the maximal reliability, $\rho_{MAX} = 0.895$. By forming different combinations of single items and their compilations, the original real-life datasets (87%) and some artificial datasets (13% of tests)—to cover the very difficult and extremely difficult tests—were used to prepare 1,296 tests with variating *N* (50–100–200), *k* (2–30), $\bar{p}$ (0.08–0.96), $df(g)$ (1–15), $df(X)$ (12–27), and α (0.74–0,98). Forming of the dataset in the derivation is described in Metsämuuronen (2020a).

Table 1 shows the essential characteristics of the tests in the derivation. Notably, the comparatively high reliabilities of the tests with difficult and extremely difficult items (0.901–0.956) reflect the fact that the artificial datasets appeared to produce notably higher item–total correlations in comparison with the real-world datasets. This matter and its effects are discussed in Section "Main limitations of the new coefficient and the process used in derivation".

These 1,296 tests produced 13,392 items with varying item characteristics (Table 2). Notably, due to the process of forming the datasets (see Appendix), the number of items with the small degrees of freedom in the item scale ($df(g) < 4$) are counted in thousands while the number of items with high degrees of freedom ($df(g) > 10$) are counted in tens.

Data analysis

The data manipulation was done in IBM SPSS 25 environment. The data mining tool, Decision Tree Analysis (DTA) and related CHAID algorithm (Kass, 1980; IBM, 2011), were used in seeking the cut-offs of the variables that explained the obvious underestimation for dimension-corrected *D*. Manual calculations were done by using a standard spreadsheet software.

Table 1. Selected characteristics of 1,296 tests used in the process

Average item difficulty ($\bar{p}$)	n. of datasets	Nature of datasets	Average ρ_{gX}	Average α_R
0 – 0.299	47	Artificial	0.867	0.901
0.3 – 0.399	112	Mainly Artificial	0.869	0.927
0.4 – 0.499	57	Mainly Real-world	0.902	0.956
0.5 – 0.599	142	Real-world	0.818	0.833
0.6 – 0.699	721	Real-world	0.821	0.867
0.7 – 0.799	217	Real-world	0.822	0.863
Total	1,296		0.830	0.873

Table 2. Selected characteristics of 13,392 items used in the process

***df*(*g*)**	**Number of items**	**Average *Rit***	**Average *Rir***	**Average *D***	**Average ρ_{BS} and ρ_{PS}**	**Average ρ_{PC} [1]**	**Number of items for ρ_{PC}**
1	7131	0.5063	0.4440	0.6284	0.6703	0.6551	2852
2	2715	0.6463	0.5658	0.6698	0.7471	0.7290	1080
3	1233	0.7266	0.6353	0.7035	0.8020	0.7781	494
4	658	0.7876	0.6818	0.7369	0.8490	0.8253	260
5	415	0.8230	0.7101	0.7535	0.8717	0.8496	169
6	335	0.8569	0.7456	0.7779	0.9044	0.8766	131
7	234	0.8832	0.7535	0.7996	0.9258	0.8959	97
8	123	0.9032	0.7697	0.8150	0.9467	0.9109	52
9	165	0.9197	0.7747	0.8363	0.9583	0.9277	62
10	140	0.9319	0.7623	0.8479	0.9667	0.9388	53
11	93	0.9427	0.7819	0.8606	0.9785	0.9471	37
12	74	0.9494	0.8062	0.8670	0.9878	0.9543	35
13–15	76	0.9488	0.7988	0.8637	0.9805	0.9537	32
	13,392						5,354

The dataset of polychoric correlation coefficient comprises 5,354 items from 518 tests by balancing the item from the real-world and artificial datasets

Principles underlying the modelling of the dimension-corrected D

Based on our knowledge of the characteristics of D and ρ_{gX}, underlying the process of deriving the correction elements, four main notes (N) were made and four consecutive principles (P) were followed:

N1. D gives a credible estimate of IDP when $df(g) = 1$ (Metsämuuronen, 2020a).
P1. D should be corrected only when $df(g) > 1$.

N2. ρ_{gX} always underestimates IDP in item analysis settings where $df(g) << df(X)$ (Metsämuuronen, 2016).
P2. The estimate by the dimension-corrected D should be higher than that by ρ_{gX} to overcome the nature of the obvious underestimation of IDP in ρ_{gX}.
N3. D tends to underestimate IDP the more the higher is $df(g)$ (Metsämuuronen, 2020a; Newson, 2002).
P3. The correction should produce more correction the higher is the $df(g)$. However, with the deterministic patterns the correction should reach the perfect value 1.

N4. In real-life settings, D reaches the maximal value 1 while ρ_{gX} does not (Metsämuuronen, 2016; 2020a; Newson, 2002).
P4. When $D = 1$, no correction is needed. Additionally, obviously, the dimension-corrected D should not exceed 1.

Because there were no theoretical reasons or empirical evidence to assume that D would under- or overestimate IDP when $df(g) = 1$ (P1), the initial model of the expected non-underestimating value for D with the linear nature is based on the assumption that there is no need to correct the estimates in the dichotomous case. Both the assumptions of linearity of the non-underestimation and that the estimate by D would be true when $df(g) = 1$ are questionable and can be debated. All in all, we do not know whether the non-underestimation should be linear or curvilinear in nature. In the deterministically discriminating dichotomous dataset with an evenly distributed score, the underestimation is elliptic

in nature (see Eq. 26 in "Potential overestimation in D_2" below). From Greiner's relation (Eq. 4) we know that, in some cases, it is a trigonometric function. Here the linear option is selected because of its simplicity.

Results

Modelling the dimension-corrected D

The dimension-corrected Somers' *D*, later called D_2, is based on modeling the underestimation in 13,392 empirical values of *D*. Figure 4 illustrates the starting point of the modeling (cf. Figure 2). The dataset suggests that the model with cubic nature $-1.02/df(g)^3 + 2.01/df(g)^2 - 1.32/df(g) + 0.95$ explains the observed distribution of *D* by $1/df(g)$ reasonably well (Figure 4). However, the model is somewhat misleading because the polynomial curve should go through the points (*1/df(g)* = 0, *D* = 1) and (*1/df(g)* = 1, *D* = 0.6284). The first point obviously indicates that, with indefinitely many categories in the item with maximal discrimination, *D* should reach the value 1 in the same manner as the other coefficients would do. The second point refers to the expectation of the level when $df(g) = 1$.

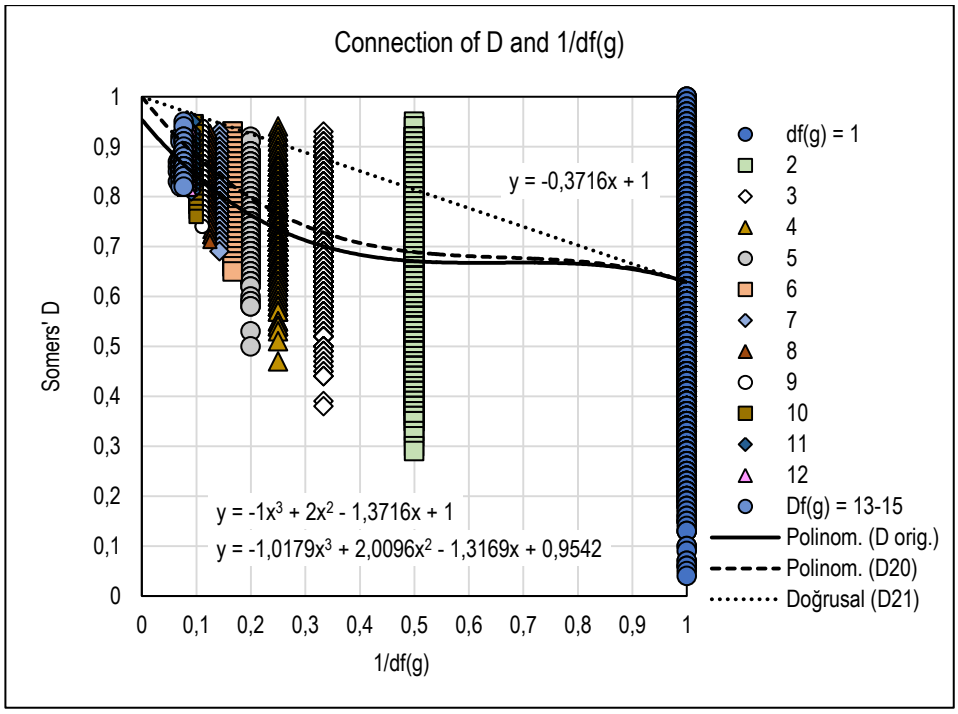

Figure 4. The original model of Somers' D and initial models D_{20} and D_{21}

The correction in *D* is based on combining the corrected third-degree model of the observed average levels of *D* against $1/df(g)$ (D_{20}, Eq. 5) and a linear model of the expected levels in varying $1/df(g)$ (D_{21}, Eq. 6). The corrected model D_{20} of third grade passing through the points (*1/df(g)* = 0, *D* = 1) and (*1/df(g)* = 1, *D* = 0.6284) is:

$$\begin{aligned} D_{20} &= 1 - \frac{1.3716}{df(g)} + \frac{2}{df(g)^2} - \frac{1}{df(g)^3} \\ &= 1 - \frac{0.3716}{df(g)} - \left(\frac{1}{df(g)} - \frac{2}{df(g)^2} + \frac{1}{df(g)^3} \right) \\ &= 1 - \frac{0.3716}{df(g)} - \frac{1}{df(g)} \left(1 - \frac{1}{df(g)} \right)^2 \end{aligned} \qquad (5)$$

where 0.3716 = 1 – 0.6284.

The magnitude of the underestimation is unknown. For the modeling purposes, the "correct" level of *D* (D_{21}) was set to be linear through the points (*1/df(g)* = 0, *D* = 1) and (*1/df(g)* = 1, *D* = 0.6284) (see Figure 4). This theoretical level of *D* in each $df(g)$ is

$$D_{21} = 1 - \frac{0.3716}{df(g)}. \quad (6)$$

The average level of discrepancy between the theoretical level and the observed level at each level of $df(g)$ is denoted by D_E:

$$\begin{aligned} D_E &= D_{21} - D_{20} \\ &= 1 - \frac{0.3716}{df(g)} - \left(1 - \frac{0.3716}{df(g)} - \frac{1}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2\right) \\ &= \frac{1}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2 \end{aligned} \quad (7)$$

and, hence, the initial correction for D is

$$D_{22} = D + D_E = D + \frac{1}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2. \quad (8)$$

The initial model D_{22} (Eq. 8) appears to be surprisingly good when it comes to increasing the average level of D. However, this model increases the magnitude of the estimates too high when D is very high in the beginning; all estimates exceeding the limits of association are of $D > 0.830$. Hence, in the second phase, a switch $(1 - D)$ related to the principle P4 was added to the correction factor D_E: $(1 - D) \times D_E$. This switch turns the correction off in the case of ultimate item discrimination when no correction is needed. An additional switch $(df(g) - 1)$ is needed to restrict the effect of $(1 - D)$ only on items with $df(g) > 1$. After these, a possible correction factor could be $(df(g) - 1) \times (1 - D) \times D_E$. The final suggestion as the dimension-corrected Somers' D is, then,

$$D_2 = D + (1 - D) \times \frac{(df(g) - 1)}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2. \quad (9)$$

By using light algebra, Eq. (9) can be further modified into

$$D_2 = 1 - (D - 1) \times (A - 1) \quad (10)$$

where D refers to Somers' $D(g|X)$ and

$$A = \frac{df(g) - 1}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2 \quad (11)$$

The correction in Eq. (10) is relevant to the positive values of D. Because of the symmetricity in Somers' D, a more general form of D_2, comprising also the negative values of D, is

$$D_2 = sign(D) \times \left(1 - \left(abs(D) - 1\right) \times (A - 1)\right), \quad (12)$$

that is, we first form the dimension correction for the absolute value of D as in Eq. (10) and then, if the original D is negative, we give the negative sign to the outcome. D_2 appears to be very potential and its characteristics are studied in what follows.

Asymptotic sampling variance and standard error of D2

Because the statistical properties of Somers' D are well documented (e.g. Agresti, 2010; Newson, 2002, 2006, 2008; Siegel & Castellan, 1988) the behavior of D_2 is known in the case of $df(g) = 1$. In the dichotomous case, the asymptotic sampling variance of D_2 can be approximated as

$$\sigma_{D_2}^2 = \sigma_D^2 = \frac{4}{D_r^4}\sum_{i,j} n_{ij}\left(D_r\left(C_{ij} - D_{ij}\right) - \left(P - Q\right)\left(N - n_i\right)\right)^2 \qquad (13)$$

that leads to asymptotic standard error

$$ASE(D_2 1) = ASE(D1) = \frac{2}{D_r^2}\sqrt{\sum_{i,j} n_{ij}\left(D_r\left(C_{ij} - D_{ij}\right) - \left(P - Q\right)\left(N - n_i\right)\right)^2} \qquad (14)$$

and, under the hypotheses of independent variables,

$$ASE(D_2 0) = ASE(D0) = \frac{2}{D_r}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}\left(P - Q\right)^2} \qquad (15)$$

where n_{ij} is the number of cases in the cell *i,j*, and n_i is the number of test-takers in the row category *i*, and

$$\begin{aligned} D_r &= N^2 - \sum_{i=1}^{r}\left(n_i^2\right) \\ C_{ij} &= \sum_{h<i}\sum_{k<j} n_{hk} + \sum_{h>i}\sum_{k>j} n_{hk} \\ D_{ij} &= \sum_{h<i}\sum_{k>j} n_{hk} + \sum_{h>i}\sum_{k<j} n_{hk} \\ P &= \sum_{i,j} n_{ij} C_{ij} \\ Q &= \sum_{i,j} n_{ij} D_{ij} \,. \end{aligned} \qquad (16)$$

Note that the formulae (13) to (16) use double than "usual" size of magnitude for *P* and *Q* seen in Eq. (3). These calculations are somewhat laborious manually. Somers (1980) offers a short-cut method found also in Siegel & Castellan (1988) and Metsämuuronen (2017b):

$$\sigma_{D2}^2 = \sigma_D^2 \approx \frac{4\left(r^2 - 1\right)\left(s + 1\right)}{9Nr^2(s - 1)} \qquad (17)$$

that leads to asymptotic standard error

$$ASE(D_2 1) = ASE(D1) \approx \sqrt{\frac{4\left(r^2 - 1\right)\left(s + 1\right)}{9Nr^2(s - 1)}} \,. \qquad (18)$$

Notably, the simplified approximation of sampling variance depends only on the *dimensions* of the variables. Hence, for all combinations of response patterns with the identical dimensions in the crosstabulation, sampling variance and related sampling error are identical.

To deriving the corresponding sampling variance for the case of $df(g) > 1$, we remember that, because of Eqs. (10), (12), and (11), after simplified,

$$VAR\left(D_2\right) = VAR\left(\text{constant} \times D\right) = \text{constant}^2 \times VAR\left(D\right). \qquad (19)$$

Then, by using the basic laws of variance, we get

$$\begin{aligned} \sigma_{D2}^2 &= VAR\left(1 - \left(D - 1\right) \times \left(A - 1\right)\right) \\ &= \left(A - 1\right)^2 \times VAR\left(D - 1\right) \\ &= \left(A - 1\right)^2 \times \sigma_D^2 \end{aligned} \qquad (20)$$

where *A* is as in Eq. (11). Then,

$$ASE(D_2 1) = \sqrt{(A-1)^2} \times ASE(D1) = \frac{2\sqrt{(A-1)^2}}{D_r^2}\sqrt{\sum_{i,j} n_{ij}\left(D_r\left(C_{ij} - D_{ij}\right) - \left(P - Q\right)\left(N - n_i\right)\right)^2} \qquad (21)$$

and, under the hypotheses of independent variables,

$$ASE(D_2 0) = \sqrt{(A-1)^2} \times ASE(D0) = \frac{2\sqrt{(A-1)^2}}{D_r} \sqrt{\sum_{i,j} n_{ij} \left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}\left(P-Q\right)^2} \quad (22)$$

and, if using the simplified short-cut by Somers (1980),

$$ASE(D_2 1) \approx \sqrt{\frac{4\left(r^2-1\right)\left(s+1\right)\left(A-1\right)^2}{9Nr^2(s-1)}} \quad (23)$$

Notably, the element $(A-1)^2 < 1$ always and, hence, the sampling variance and standard error of the estimates by D_2 are always smaller than those by Somers' D. When testing the null hypothesis $H_0 : D_2 = 0$ (which is usually not a relevant option in the item analysis settings though), we can use the statistic

$$z = \frac{D_2}{ASE(D_2 0)}. \quad (24)$$

This value is approximately normally distributed with mean 0 and standard deviation 1 when the null hypothesis is true.

A numerical example of D2

As a numerical example of calculating D_2, assume a simple polytomous dataset with N = 25 cases as in Table 3 adapted from Cox (1974, p. 177) and Drasgow (1986, p. 70). Let us assume that the dataset would concern an item g and the score X.

Table 3. A hypothetic dataset (Cox, 1974; Drasgow, 1986)

g	***X***	***g***	***X***	***g***	***X***	***g***	***X***	***g***	***X***
0	72	1	77	1	87	1	99	2	85
0	88	1	78	1	88	1	101	2	96
0	112	1	80	1	92	1	104	2	96
1	69	1	81	1	92	1	104	2	103
1	72	1	86	1	93	1	108	2	104

Used by permisson of Biometric society

Table 4. Contingency table based in Table 3

		X																		
		69	72	77	78	80	81	85	86	87	88	92	93	96	99	101	103	104	108	112	SUM (g_i)
g	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	3
	1	1	1	1	1	1	1	0	1	1	1	2	1	0	1	1	0	2	1	0	17
	2	0	0	0	0	0	0	1	0	0	0		0	2	0	0	1	1	0	0	5
	SUM (X_j)	1	2	1	1	1	1	1	1	1	2	2	1	2	1	1	1	3	1	1	25

In the first phase, Somers' D is calculated. For this, a cross-table is formed (Table 4). For the manual calculation of Somers' D, the sums of concordant pairs (P) and discordant pairs (Q) are formed (see Siegel & Castellan, 1988; Metsämuuronen, 2017b). For these, the cell frequencies are denoted by n_{ij}. For the concordant pairs, we calculate how many cases are there in the cells below and to the *right* of the cell n_{ij}. These are denoted by N_{ij}^{+}. Correspondingly, the discordant pairs denoted by N_{ij}^{-} are found in the cells below and to the *left* of the cell n_{ij}. All possible values for N_{ij}^{+} and N_{ij}^{-} are computed and these are multiplied by the related n_{ij}. The number of all the pairs in the same direction is

$$P = \sum_{ij} n_{ij} N_{ij}^{+} = 1\times 20 + 1\times 12 + 6\times 5 + 6\times 4 + 2\times 2 = 90.$$

Correspondingly, the number of the pairs in the opposite direction is

$$Q = \sum_{ij} n_{ij} N_{ij}^{-} = 1\times 1 + 1\times 9 + 1\times 22 + 6\times 1 + 2\times 3 + 2\times 4 + 1\times 5 = 57.$$

By using Eq. (3), the estimate of the association by Somers' *D*("score dependent"), that is, *D*("*g* in condition of *X*")‡ is

$$\hat{D}(g|X) = \frac{2(P-Q)}{N^2 - \sum_{j=1}^{g}(n_{gj}{}^2)} = \frac{2\times(90-57)}{25^2-(3^2+17^2+5^2)} = \frac{66}{302} = 0.219\cdot$$

For the dimension correction, we need the correction factor *A* (Eq. 11). With three categories in the item scale, *df*(*g*) = 2 and, hence,

$$A = \frac{df(g)-1}{df(g)}\left(1-\frac{1}{df(g)}\right)^2 = \frac{1}{2}\times\frac{1}{4} = 0.125\cdot$$

Because of Eq. (10), the estimate of the observed association of the item and score by D_2 is

$$\hat{D}_2 = 1-(D-1)\times(A-1) = 1-(0.219-1)\times(0.125-1) = 0.317$$

with standard error

$$ASE(\hat{D}_2 1) = \sqrt{(A-1)^2}\times ASE(D1) = 0.875\times 0.242 = 0.212\,.$$

As benchmarks, the estimates of the observed association based on the mechanics of Pearson's product-moment correlation are $\hat{\rho}_{gX} = 0.185$ and, after corrected for the inflation, $\hat{\rho}_{gP} = 0.139$. The estimate of the inferred association by polyserial correlation is $\hat{\rho}_{PS} = 0.216$ and the corresponding estimate by the polychoric correlation is $\hat{\rho}_{PC} = 0.123$ though the last value depends of the estimation method in some extent.

General characteristics of D_2

D_2 behaves according to the four principles set for the correction. First, the estimates by *D* are not corrected when *df*(*g*) = 1. Second, the estimates by D_2 tend to be, generally, higher than those by ρ_{gX}, ρ_{gP}, and ρ_{PC}, and close to those by ρ_{PS}, although without the obvious overestimation (see Figures 5 and 6). Third, the higher is *df*(*g*) the greater the correction is in D_2. Fourth, D_2 does not correct *D* when item discrimination is deterministic and *D* = 1. Of the 13,392 items on the simulation, none showed a value that was out of range regarding the limits of correlation.

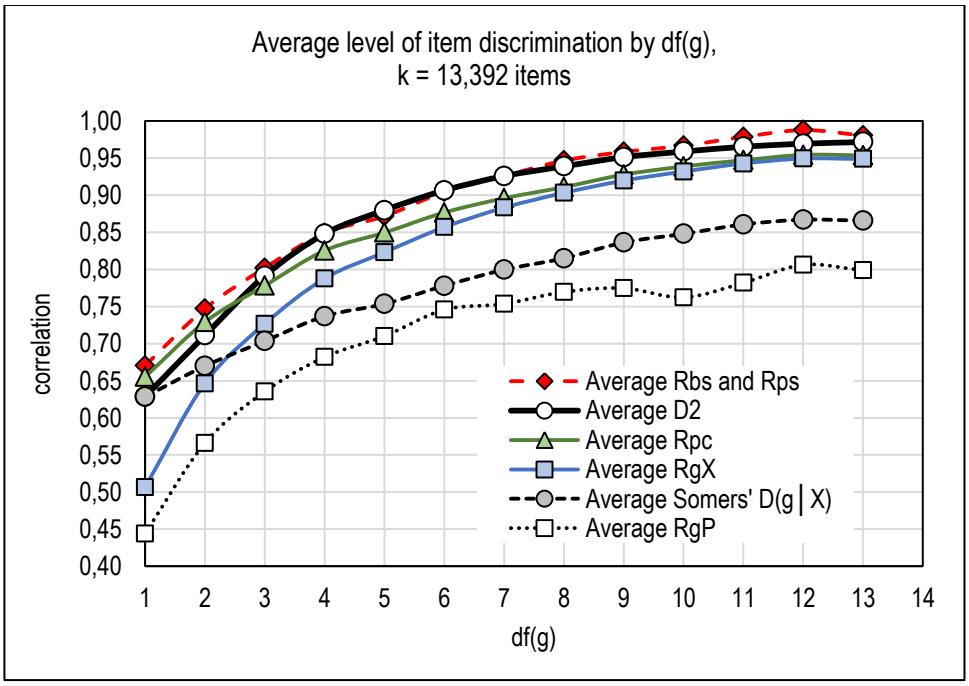

Figure 5. Average estimates of selected indices of IDP by varying df(g)

‡ Again, it is worth noting the specific wording when it comes to textbooks and outputs related to Somers' *D*. All the generally known textbooks and software packages use the term "score dependent" for this formula. However, it tells us how well the item discriminates the test-takers after they are ordered by the score, that is, the order in the item depends on the order in the score.

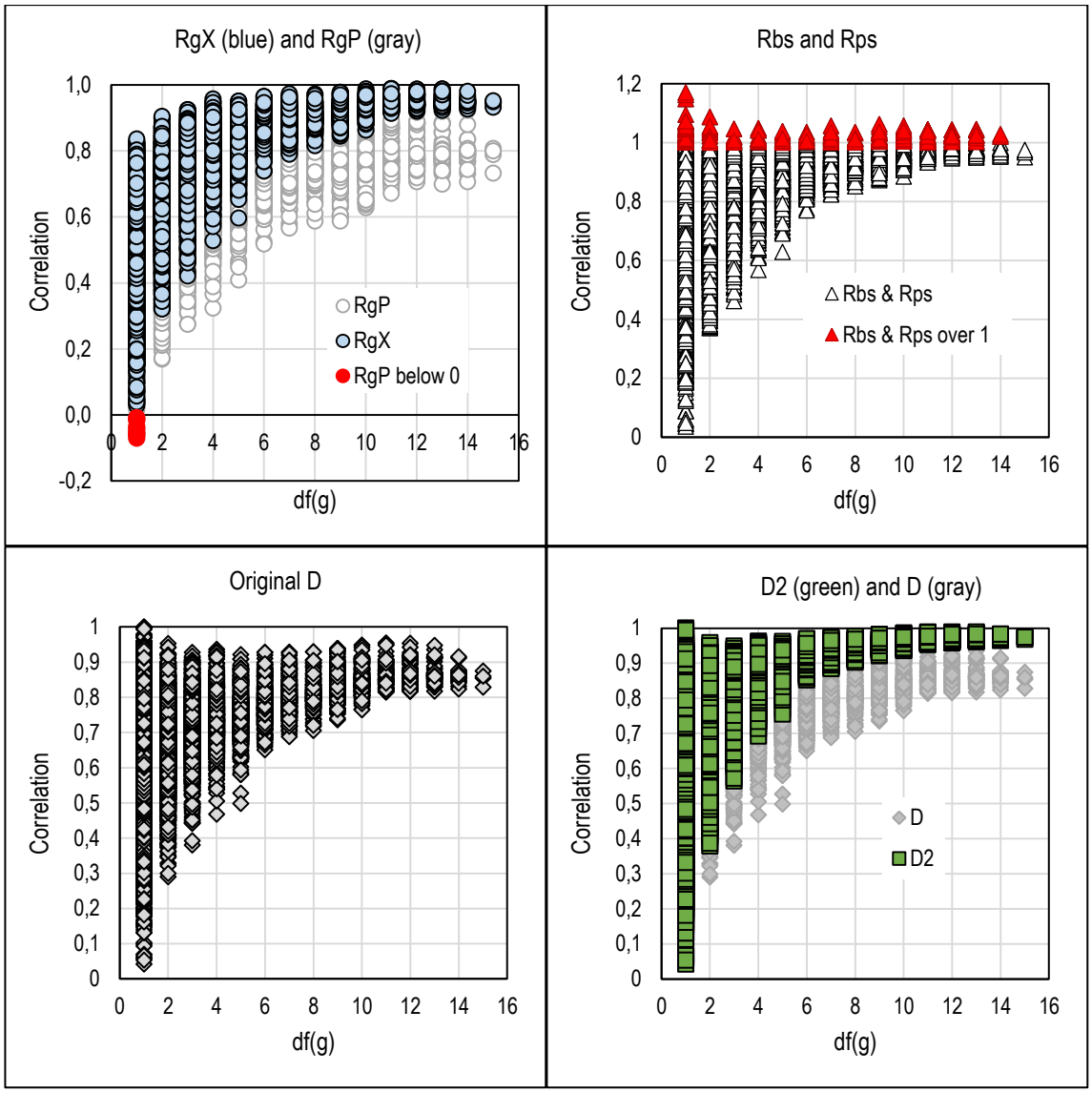

Figure 6. Distributions of the estimates by ρ_{gX} (RgX), ρ_{gP} (RgP), ρ_{BS} (R_{BS}), ρ_{PS} (R_{PS}), original D, and D_2 by varying df(g)

Overall, when it comes to correcting the underestimation in D, D_2 behave logically at all levels of $df(g)$ used in the simulation. On average, D_2 underestimates the IDP remarkably less than Somers' D, and notably less than ρ_{gX} and ρ_{PC} as was the motivation for the derivation. We may also note that the average magnitudes of the estimates by D_2 tend to follow those of ρ_{PC} when $df(g) < 3$—as was the case in the numerical example with Table 3. Notably, however, in the simulation dataset, ρ_{PC} tends to start to follow the magnitude of ρ_{gX} when $df(g) > 6$. This indicates that ρ_{PC} tends to start to underestimate the IDP the same way ρ_{gX} does with high degrees of freedom in the item. To some extent, the average magnitudes of the estimates by D_2 tends to follow those by ρ_{PS}, although without the obvious overestimation (see Figure 6). Notably, the variability in the magnitudes of the estimates by D_2 is smaller than that by D at each level of $df(g) > 1$ (Figure 6).

Limits of D2

When $df(g) = 1$,

$$A = \frac{df(g)-1}{df(g)}\left(1-\frac{1}{df(g)}\right)^2 = \frac{0}{1}\left(1-\frac{1}{1}\right)^2 = 0$$

and, then,

$$D_2 = 1-(D-1)\times(A-1) = D.$$

When $D = 1$, regardless $df(g)$,

$$D_2 = 1-(D-1)\times(A-1) = 1-0\times(A-1) = 1.$$

The limit of degrees of freedom related to D_2 reflects the peculiarity of the magnitudes of the estimates of IDP as discussed above. It is worth noting that the dimension-corrected coefficient is created for the case that the degrees of freedoms of two variables are far from each other. In the theoretical extreme case when $df(g) = \infty$, that is, with the continuous items and infinite number of test takers with different item score (to form infinite number of categories in the item),

$$\lim A = \lim \frac{df(g)-1}{df(g)}\left(1-\frac{1}{df(g)}\right)^2 = 1\times 1^2 = 1 \qquad (25)$$

and, then, the correction in Eq. (10) leads us to a triviality that $D_2 = 1-(D-1)\times(A-1) = 1$ seemingly regardless the actual association between the item and the score. However, the indefinitely long "parallel tests" approximates the ultimate magnitude of $\rho_{gX} = D_2 = 1$. Hence, within the item analysis settings, with the indefinitely many categories in the item(s), the score would contain also indefinite number of categories and, then, D approximates the magnitude of 1. However, Eq. (25) hints that when two continuous variables with different scales are *independent* from each other, another kind of correction than Eqs. (10) and (12) may be needed. *This restriction of D_2 is necessary to keep in mind if applying it to items with continuous scale with infinite number of categories.* However, we may remember, that the continuous scale itself alone does not lead to triviality of $D_2 = 1$ because, even with the continuous values in the scales, the number of categories in the item may be small and then, obviously, $df(g) << \infty$. This matter is relevant in relation to the measurement modelling settings where the items may be weighted by a factor loading. Regardless the seemingly continuous scale, the actual weighting of, for example, binary items leads to two categories; now instead of categories 0 and 1, we may have categories 0 and 0.678, as an example. Another viewpoint to this restriction of using D_2 is that the contemporary procedures related to item analysis are usually related to non-continuous scales in the item. Hence, the condition of $df(g) = \infty$ is a highly theoretical option and does not relate with the real-life item analysis settings as we face those today.

Obvious underestimation in D_2

A simple criterion for the obvious underestimation in the estimates by D_2 is whether the magnitudes of the estimates are lower than those by ρ_{gX}. Knowing that the estimate by ρ_{gX} is practically always an underestimate for IDP, lower values would strictly be indicative of even *more* underestimation in D_2. Of the 7,131 items on the simulation with $df(g) = 1$, the original D (= D_2) included 12 cases (0.1%) where $\rho_{gX} > D$. All these cases came from the artificial datasets with relatively high value of ρ_{gX} (see Table 5).

Table 5. Selected characteristics of the factors explaining the obvious underestimation in the different options for dimension-corrected D

Factor	Cut-off[1]	$\rho_{gX} > D_2$ (n = 36)	$\rho_{gX} > D_2$ (%)
Type of the dataset	artificial	36	100
ρ_{gX}	> 0.853	20	55.6
D	> 0.830	16	44.4
Number of items in the test (k)	2 or 3	13	36.1
$df(g)$	9–13	11	30.6

The groups and cut-offs were suggested by Decision Tree Analysis (DTA; IBM, 2011). Each factor was analyzed individually by using CHAID algorithm (Kass, 1980) without further restrictions.

When $df(g) > 1$, additionally, we find 36 additional estimates (0.3%) by D_2, where the magnitude of the estimate by ρ_{gX} is higher than that by D_2. All these obvious underestimates by D_2 (0.4% of the estimates) come from the artificial dataset with an artificial combination of high item discrimination and low item difficulty. As a benchmark, with the original Somers' D when $df(g) > 1$, as many as 62% of the estimates in the simulation datasets are obviously underestimated. Hence, the number of the clearly underestimated estimates by D_2 seems relatively low. Some of the characteristics of the obvious underestimates are collected in Table 5. It seems that the probability of obtaining obvious underestimation in real-life datasets is very low when using D_2.

Potential overestimation in D_2

If the magnitude of the estimates by D_2 would be higher than 1, those would be obvious overestimates. In the simulation, none of the items showed this behavior. Otherwise, possible overestimation is not easy to evaluate in strict terms when using real-world datasets. One potential criterion for the overestimation in these cases is the theoretical,

maximally discriminating Guttman-patterned datasets (Guttman, 1950). In the Guttman pattern, with $df(g) = 1$, D gives the maximal estimate 1 while the estimates by ρ_{gX} are always smaller than 1; the maximal ρ_{gX} is reached when $p = 0.5$. Assuming a score without ties, the highest value of item–total correlation approximates $\rho_{gX}^{\max} = 0.866$ (see Metsämuuronen, 2016) and, hence, the lowest point of the difference is $D - \rho_{gX} = 1 - 0.866 = 0.134$. This boundary is illustrated in Figure 7.

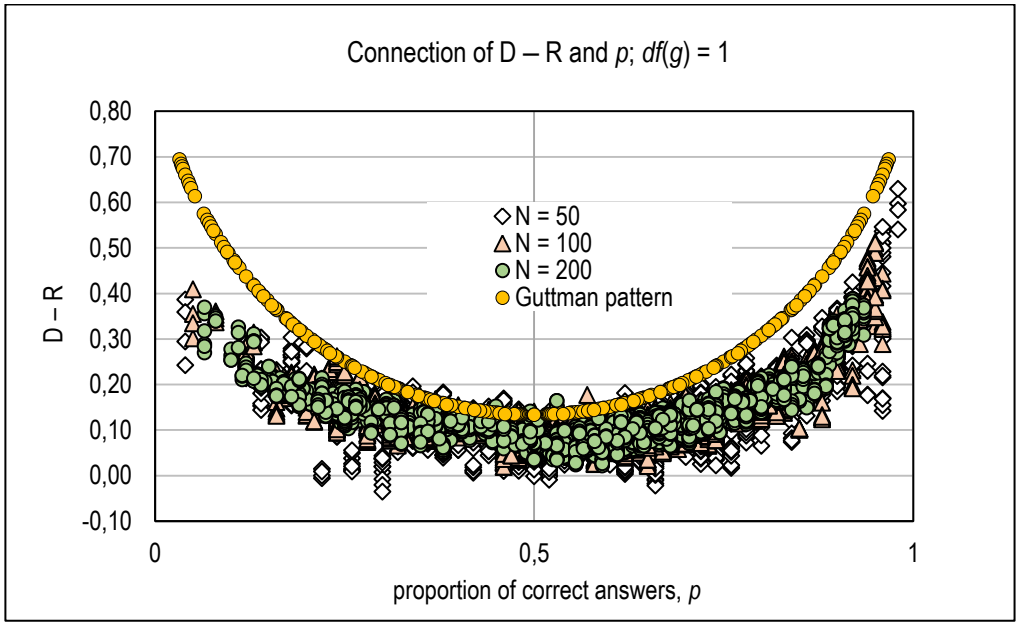

Figure 7. Guttman-pattern as a limit for the possible overestimation

In the binary case, Guttman boundary follows an ellipse with the parameters $x_0 = 0.5$, $y_0 = 0$, $a = 0.5$ and $b = \rho_{gX}^{\max} = 0.866$:

$$\frac{(X - x_0)^2}{a^2} + \frac{(Y - y_0)^2}{b^2} = \frac{(p - 0.5)^2}{0.5^2} + \frac{(\rho_{gX} - 0)^2}{0.866^2} = 1 \qquad (26)$$

where p is the item difficulty and 0.866 refers to the limit of the maximum value of Pearson correlation in the deterministic pattern in the dataset. From (26) we solve ρ_{gX} :

$$\rho_{gX} = \sqrt{\left(1 - \frac{(p - 0.5)^2}{0.5^2}\right) \times 0.866^2} \qquad (27)$$

and, then, in Guttman-patterned items,

$$D - \rho_{gX} = 1 - \sqrt{\left(1 - \frac{(p - 0.5)^2}{0.5^2}\right) \times 0.866^2} \qquad (28)$$

This model is used as a rough tool to evaluate the possible overestimation in D_2 (Figure 8). In the real-world datasets in the simulation, 18 out of 13,392 estimates by Somers' D (0.13%) exceeded this limit, and, in the artificial datasets, 33 (0.25% of all items). In all these cases, the magnitude of the overestimation is nominal (near zero units of correlation). Notably, in comparison with the original D, D_2 produced only one additional estimate with non-significant magnitude that exceeded the boundary of the Guttman pattern.

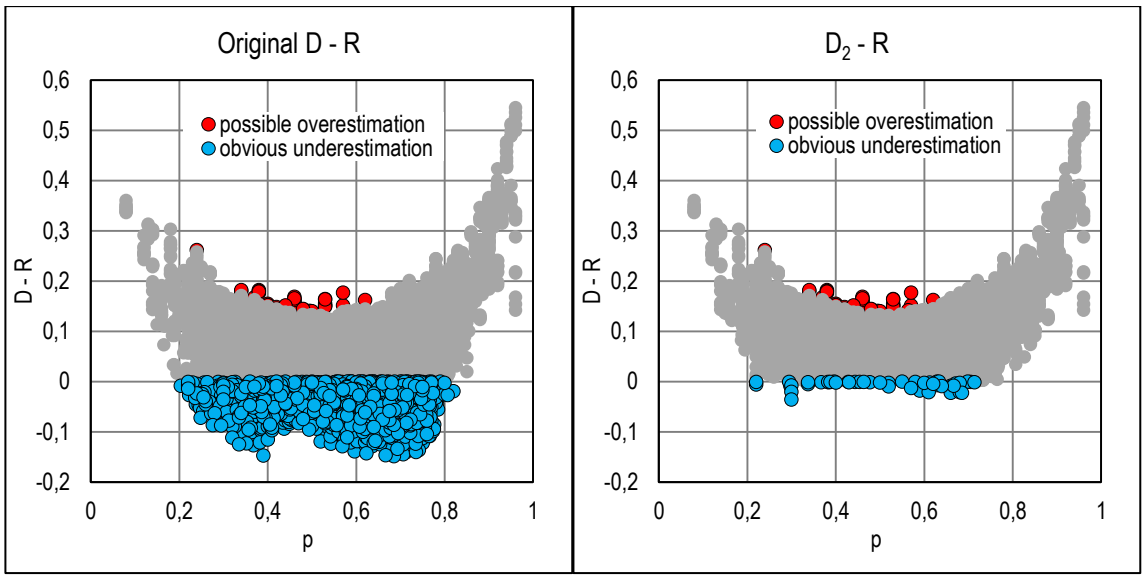

Figure 8. Possible overestimation and obvious underestimation in D and D_2

Conclusions

A dataset of 13,392 real-life items with varying characteristics was used to model the underestimation in *D* and to derive the "dimension-corrected Somers' *D*" for the measurement modelling settings. In its general form, the new coefficient is $D_2 = sign(D)\times\left(1-\left(abs\left(D\right)-1\right)\times\left(A-1\right)\right)$ where *D* is the uncorrected Somers' $D\left(g\middle|X\right)$ (i.e. "item in condition of score" or "score dependent" in the standard outputs of the software packages) and $A=\frac{df\left(g\right)-1}{df\left(g\right)}\left(1-\frac{1}{df\left(g\right)}\right)^2$ where *df*(*g*) is the number of marginal categories in the item minus 1. Within the normal range of non-pathological item discrimination, that is, with positive association between the item and score, $D_2 = 1-\left(D-1\right)\times\left(A-1\right)$.

D_2 equals Somers' *D* in two cases: when *df*(*g*) = 1, that is, in binary datasets, and when $D = \pm 1$, that is, with deterministic item discrimination. As do all the classical estimators of IDP, D_2 approaches the value $D_2 = 1$ when the number of categories in the item scale approximates the scale of the score. Additionally, in a highly theoretical case of infinite number of categories in the item (and, consequently, in the score), D_2 approximates $D_2 = 1$ seemingly regardless the actual value of Somers' *D*. Under this condition, however, also D_2 (as well as all estimators of IDP because of the mechanical connection between the items and the score) approximates 1.

In the datasets in the simulation, D_2 showed very few cases of obvious underestimation and overestimation. The correction is simple but seems to get an effective result. With certain restrictions discussed in the section "Main limitations of the new coefficient and the process used in derivation", D_2 seems to be superior over other indices in comparison not only in binary cases but also in cases where the degrees of freedom increase up to 15 categories; more categories were not used in the simulation.

Overall, D_2 corrects the underestimation in *D* effectively and hence, in most cases, the magnitude of the estimates expectedly draws us nearer the real IDP that those by ρ_{gX}. The number of obvious cases of underestimation by D_2 is reduced remarkably in comparison to the original Somers' *D*—from 62% to 0.3% of the estimates with *df*(*g*) > 1. . In most of these obvious underestimations, the magnitude was close to zero units of the correlation. The number of estimates with a possible overestimation did not increase when the boundary of the deterministically discriminating Guttman pattern was kept as the criterion. The possible overestimation in the dimension-corrected *D* may need more studies though. Other limitations of the new coefficient are discussed in the section "Main limitations of the new coefficient and the process used in derivation".

Discussion

Some advantages of D_2

Combining the advantages of Somers' *D* from Metsämuuronen (2020a) and Newson (2002) as well as the empirical findings in this article, the dimension-corrected Somers' *D* could be proposed as one of the "superior alternatives" to ρ_{gX} and ρ_{gP} and, in some extent, also to ρ_{BS}, ρ_{PS}, and ρ_{PC} in reflecting the IDP in item analysis settings because of the

following reasons:

1. D_2 reaches the values +1 and −1 accurately while ρ_{gX} and ρ_{gP} cannot reach the limits within practical measurement modeling settings, ρ_{BS} and ρ_{PS} may easily give obvious overestimates and underestimates, and ρ_{PC} cannot reach the extreme value with standard procedures. Additionally, because the magnitude of the estimates by ρ_{PC} tend to follow those by ρ_{gX}, the estimates seems to underestimate IDP when $df(g) > 7$.
2. D_2 is more robust for extreme observations and for nonlinearity than ρ_{gX} and ρ_{gP}.
3. D_2 is superior to ρ_{gX} and ρ_{gP} and to some extent also to ρ_{PC} with both dichotomous and polytomous items, because it is highly probable that D_2 produces an estimate that underestimates the IDP less than ρ_{gX}, ρ_{gP}, and ρ_{PC} do.
4. D_2 does not produce out-of-range values as do ρ_{BS} and ρ_{PS}.
5. D_2 utilizes the known composite of items in the analysis that is easy to use in further research while ρ_{PC} refers to an unknown, unreachable, and hypothetical composite that is difficult to use.
6. D_2 is applicable and accurate with large, small, non-normal, or sparse cross-tables while the applicability and accuracy of the estimation result of the ρ_{PC} depends on the form of cross-tabulation and normality of the phenomenon.
7. D_2 has a logical directional nature from the modern measurement–modeling viewpoint; while D_2 indicates how well the latent trait (score) explains the behavior in the manifested variable (item), the other estimators in comparison (ρ_{gX}, ρ_{gP}, ρ_{BS}, ρ_{PS}, and ρ_{PC}) tell us about the unspecified association of the variables.
8. D_2 increases the possibilities of detecting the maximally discriminating test items in comparison with ρ_{gX}, ρ_{gP}, ρ_{BS}, ρ_{PS}, and ρ_{PC}. These kinds of datasets where the order of the test-takers in the item is the same as in the score are more frequent with small datasets relevant in, for example, classroom testing settings. In these patterns, unlike the other estimators, $D_2 = 1$ always irrespective of the number of cases, degrees of freedom of the item and the score, the number of tied values, difficulty levels in the items, or the number of items on the test.
9. D_2 is reasonably easy to calculate even manually in practical test settings such as classroom testing, while calculation of ρ_{PC} requires specific software packages and complex procedures.

Main limitations of the new coefficient and the process used in derivation

One obvious challenge in generalizing the new coefficient is that D_2 is *developed for item analysis settings*. In these settings, always $df(g) << df(X)$, and the items and the score are mechanically connected. Notably, the dimension-correction leads, automatically, to approximate the perfect value $D_2 = 1$ (or, in the ultimate pathological case, to $D_2 = -1$) when the item is a continuous one and the sample size is large. Because of this, the applicability of D_2 may be reduced outside the measurement modeling settings. Hence, it is not wise to use D_2 as a general coefficient without further studies and possible amendments. The coefficient is suitable for the negative values of D though, however, these are pathological cases in item analysis settings.

Second, during the process, the *benchmark of the possible underestimation was the Pearson's product–moment correlation* coefficient while, perhaps, some other coefficient would have been more appropriate. Anyhow, the correction seems to bring us nearer the true IDP also in comparison with other indices. More studies are needed in this respect. Specifically, from this viewpoint, an interesting benchmark would be a coefficient called *r*-polyreg correlation, that is, an *r*-polyserial estimated by regression correlation (Lewis et al., 2003 cited by Livinstone & Dorans, 2004). This coefficient, developed to overcome the challenge of obvious overestimation in ρ_{BS} and ρ_{PS}, can be used with binary or polytomously scored items and it produces estimates that do not exceed 1, nor does it rely on bivariate normality assumptions (Moses, 2017).

Third, the correction elements in D_2 are based on simulation with empirical items that embed the *limitations of the original datasets* to a certain extent. We do not know how much the estimates depend on the original dataset. However, we note that there are no numerical sub-coefficients in the correction factors in Eqs. (10) and (11). Hence, to some extent, the new coefficient is free from the original dataset and the correction is more general than is the case when it includes specific numerical coefficient(s) strictly dependent on the underlying dataset. Seeing that the values arrived at are based on 13,392 items with varied characteristics and a strong base in the real world, the estimates are likely to be quite stable in relation to real-life settings of testing, although wider simulations may give more insights in the matter. Cross-validating the model by using datasets from the same basic population and same test items would not challenge the models profoundly. Specifically, such simulation where the degrees of freedom of the item are higher than seven

would enrich our knowledge of the coefficient; the dataset used in the simulation in this article contained few items of these kind. Also, simulations regarding the possible over and underestimation of association in general would benefit us.

The fourth limitation is that *dimension–correction is modeled for Somers'* $D(g|X)$ and not for Somers' $D(X|g)$ or for symmetric *D.* Hence, the correction cannot necessarily be generalized though it may carry general elements for *df*(*X*) >> *df*(*Y*) or *df*(*Y*) >> *df*(*X*). From the measurement–modeling viewpoint, however, the direction $D(g|X)$ ("item in condition of score") is more relevant than $D(X|g)$ ("score in condition of item"). In any case, generally, it would be valuable to study whether the correction elements developed in this study are valid also in the symmetric case and in the case of $D(X|g)$. It may appear that when the degrees of freedoms of the variables are nearer each other we may need the degrees of freedom of *both* variables in the correction—now, only *df*(*g*) appeared to be significant factor in the correction.

Some suggestions for the further studies

One natural direction for the further studies is to study the new coefficient itself. First, larger simulations would confirm the characteristics of the new coefficient. While there is a need for a simulation with higher degrees of freedom than seven to see how much a small number of estimates affected the correction elements at this range, simulations are also needed to confirm or alter the coefficient in case the degrees of freedoms are close to each other.

Second, being a new index of correlation related to item discrimination, it would be valuable to compare the characteristics of the new coefficient with some other, new, well-behaving coefficients, such as *r*-polyreg correlation.

Third, being a new coefficient of association, its properties may be valuable to study from that viewpoint as well. We may also ask: does the coefficient carry the essential characteristics of Somers' *D* at all or should it be taken as a totally new coefficient based on Somers' *D*?

Fourth direction for future research is to study the new coefficient in relation with other relevant aspects of measurement modeling. Then, the new coefficient may have relevance when estimating "dimension-corrected reliability" of the test score, for example. Item–total correlation, which always underestimates the connection of the score and the item, is embedded in all widely used estimators of reliability because, in the classical forms of reliability, the element σ_X^2 can be expressed by using ρ_{gX} : $\sigma_X^2 = \left(\sum_{g=1}^{k} \sigma_g \rho_{gX} \right)^2$ (Lord & Novick, 1968), where *k* refers to the number of items, booklets, or partitions of the test items. This matter concerns such classical estimators of reliability as Spearman–Brown prophesy formula (Brown, 1910; Spearman, 1910), Flanagan or Flanagan–Rulon formula (Flanagan, 1937; Rulon, 1939), the family of Guttman's Lambda (Guttman, 1945) as well as the classical formula KR20 by Kuder and Richardson (Kuder & Richardson, 1937), and its generalized version coefficient alpha (timewise Guttman, 1945; Gulliksen, 1950; Cronbach, 1951). As the magnitude of ρ_{gX} is always lower than it should be, Metsämuuronen (2016) argued for that this mechanical underestimation is at least one of the reasons why the classical coefficients tend to underestimate reliability. We may note that item–total correlation is embedded also in the processes of calculating more advanced estimators of reliability based on factor analysis such as McDonald's Omega (McDonald, 1999) and maximal reliability (e.g. Li, 1997; Raykov 2004; 2005 onwards) because factor loadings in orthogonal rotation are (Pearson) correlations between the (weighted) items and the (latent) factor. This means that the very essence of factor loading is item–scale correlation. Perhaps D_2 could be used instead of Pearson correlation (or some other estimator) in these formulae and procedures. This may lead us to correct the estimates obtained by the classical estimators such as coefficient alpha and maximal reliability and, hence, we can get nearer the *real* reliability than we can by using the traditional estimators or at least this can give us the "dimension-corrected reliability".

Fifth, the directional nature of the coefficient and its possible usefulness within the modern measurement modeling processes may be worth studying. The nondirectional Pearson product–moment correlation coefficient and the family of polychoric correlations are deeply set in the procedures in EFA and SEM analyses. A relevant underlying question that arises from the directional D_2 and the underlying Somers' *D* is why in the first place are we willing to use the *non*directional correlation coefficients in our testing and measurement modeling settings while the whole philosophy of measurement modeling is based on the idea of directionality the latent trait manifest as the score or the measurement scale determines the observed behavior and not the other way round (e.g. Byrne, 2001; Metsämuuronen, 2017b): in psychometric theory, the overall trait being measured generally drives examinees' responses to, and, thus, scores/measurement scales on individual items (see the discussion in Metsämuuronen, 2020a). Then, the family of the directional coefficients of correlation seems to be at least possible if not suggestible alternatives for measurement modeling. The directional, dimension-corrected correlation coefficient D_2 could be a relevant option to consider from this point of view.

Overall, Somers' *D* seems to be a very potential tool within measurement modeling settings because of its natural characteristic of directing the connection of two variables the same way as we find in the settings of structural equation modeling. With dimension-correction, D_2 could be an even *more* useful tool in both item analysis settings and in measurement modeling. It may help us get closer to the *real* connection between the latent and manifest variables, *real* item discrimination, and *real* reliability.

Acknowledgements

Sincere thanks to Dr. Roger Newson, *research associate* at the Faculty of Medicine, School of Public Health, Imperial College London, for suggesting Greiner's relation to understand the underestimation in Somers' *D*; he also led to other useful resources on the topic.

References

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed). Wiley.

Aslan, S., & Aybek, B. (2020). Testing the effectiveness of interdisciplinary curriculum-based multicultural education on tolerance and critical thinking skill. *International Journal of Educational Methodology, 6*(1), 43–55. https://doi.org/10.12973/ijem.6.1.43.

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

Byrne, B. M. (2001). *Structural equation modeling with AMOS. Basic concepts, applications, and programming.* Lawrence Erlbaum Associates, Publishers.

Cox, N. R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics, 30*(1), 171–178. https://doi.org/10.2307/2529626.

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

Cureton, E. E. (1956). Rank–biserial correlation. *Psychometrika*, *21*(3), 287–290. https://doi.org/10.1007%2FBF02289138.

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

Delil, A., & Ozcan, B.N.(2019). How 8th graders are assessed through tests by mathematics teachers? *International Journal of Educational Methodology, 5*(3), 479–488. https://doi.org/10.12973/ijem.5.3.479.

Drasgow, F. (1986). Polychoric and polyserial correlations. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences- Vol 7* (pp. 68–74). John Wiley.

Educational Testing Service (2020). Glossary of standardized testing terms. Educational Testing Service. https://www.ets.org/understanding_testing/glossary/

Finnish Education Evaluation Centre (2018). *National assessment of learning outcomes in mathematics at grade 9 in 2002* (Unpublished dataset opened for the re-analysis 18.2.2018). Finnish National Education Evaluation Centre.

Flanagan J. C. (1937). A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology*, *28*(1), 17–21. https://doi.org/10.1037/h0057430.

Goktas, A., & Isci. O. A. (2011). Comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. *Methodological Notebooks/ Metodoloski Zvezki*, *8*(1), 17–37.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*(268), 732–764. https://doi.org/10.1080/01621459.1954.10501231.

Greiner, R. (1909). Uber das fehlersystem der kollektivmaßlehre [Of the error systemic of collectives]. *Journal of Mathematics and Physics/ Zeitschift fur Mathematik und Physik*, *57*, 121–158, 225–260, 337–373.

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

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

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfield, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction.* Princeton University Press.

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

Howard K. I, & Forehand, G. A. (1962). A method for correcting item-total correlations for the effect of relevant item inclusion. *Educational and Psychological Measurement, 22*(4), 731–735. https://doi.org/10.1177/001316446202200407.

IBM (2011). *IBM SPSS Decision trees 20.* ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Decision_Trees.pdf

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*(2), 119–127. https://doi.org/10.2307/2986296.

Kendall, M. (*1938). A new measure of rank correlation. Biometrika*, *30*(1/2), 81–93. https://doi.org/10.2307/2332226.

Kendall, M. (1949). Rank and product–moment correlation. *Biometrika, 36*(1/2), 177-193. https://doi.org/10.2307/2332540.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391.

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

Liu, F. (2008). *Comparison of several popular discrimination indices based on different criteria and their application in item analysis*. University of Georgia.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison–Wesley Publishing Company.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*(6), 921–943. https://doi.org/10.1177/0013164402238082.

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

Metsämuuronen, J. (2016). Item–total correlation as the cause for the underestimation of the alpha estimate for the reliability of the scale. *GJRA - Global Journal for Research Analysis*, *5*(1), 471–477.

Metsämuuronen, J. (2017a). *Essentials of research methods in human sciences. Vol 1: Elementary basics.* SAGE Publications.

Metsämuuronen, J. (2017b). *Essentials of research methods in human sciences. Vol 3: Advanced analysis.* SAGE Publications.

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

Metsämuuronen, J. (2020b). Generalized discrimination index. *International Journal of Educational Methodology*, *6*(2), 237-257. https://doi.org/10.12973/ijem.6.2.237

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

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and Median differences. *The Stata Journal, 2*(1), 45–64.

Newson, R. (2006). Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal, 6*(3), 309–334.

Newton, R. (2008). *Identity of Somers' D and the rank biserial correlation coefficient.* Roger Newson http://www.rogernewsonresources.org.uk/miscdocs/ranksum1.pdf

Pearson, K. (1896). Mathematical contributions to the theory of evolution III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical 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 Transactions of the Royal Society A. Mathematical, Physical and Engineering Sciences, 195*(262–273), 1–47. https://doi.org/10.1098/rsta.1900.0022

Pearson, K. (1903). I. Mathematical contributions to the theory of evolution. —XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society A. Mathematical, Physical and Engineering Sciences, 200*(321–330), 1–66. https://doi.org/10.1098/rsta.1903.0001

Pearson, K. (1905). *On the general theory of skew correlation and non-linear regression*. Dulau & Co.

Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika*, *9*(1–2), 116–139. https://doi.org/10.1093/biomet/9.1-2.116

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modeling approach. *British Journal of Mathematical and Statistical Psychology, 57*(1), 21–27. http://doi.org/10.1348/000711004849295

Raykov, T. (2005). Studying group and time invariance in maximal reliability for multiple-component measuring instruments via covariance structure modeling. British *Journal of Mathematical and Statistical Psychology, 58*(Pt 2), 301–317. http://doi.org/10.1348/000711005X38591

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

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill.

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

Spearman, C. (1910). Correlation computed with faulty data. *British Journal of Psychology*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Van der Ark, L. A., & Van Aert, R. C. M. (2015). Comparing confidence intervals for Goodman and Kruskal's gamma coefficient. *Journal of Statistical Computation and Simulation, 85*(12), 2491–2505. https://doi.org/10.1080/00949655.2014.932791

Wendt, H. W. (1972). Dealing with a common problem in social science: A simplified rank biserial coefficient of correlation based on the U statistic. *European Journal of Social Psychology*, *2*(4), 463–465. https://doi.org/10.1002/ejsp.2420020412

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