Research Article doi: 10.12973/ijem.6.1.207

International Journal of Educational Methodology
Volume 6, Issue 1, 207 - 221.
ISSN: 2469-9632
http://www.ijem.com/

Somers' D as an Alternative for the Item–Test and Item-Rest Correlation Coefficients in the Educational Measurement Settings

Jari Metsämuuronen*
Finnish Education Evaluation Centre,
FINLAND

NLA University College,
NORWAY

Received: November 1, 2019 ▪ Revised: January 28, 2020 ▪ Accepted: February 13, 2020

Abstract: Pearson product–moment correlation coefficient between item *g* and test score *X*, known as item–test or item–total correlation (*Rit*), and item–rest correlation (*Rir*) are two of the most used classical estimators for item discrimination power (IDP). Both *Rit* and *Rir* underestimate IDP caused by the mismatch of the scales of the item and the score. Underestimation of IDP may be drastic when the difficulty level of the item is extreme. Based on a simulation, in a binary dataset, a good alternative for *Rit* and *Rir* could be the Somers' *D*: it reaches the ultimate values +1 and –1, it underestimates IDP remarkably less than *Rit* and *Rir*, and, being a robust statistic, it is more stable against the changes in the data structure. Somers' *D* has, however, one major disadvantage in a polytomous case: it tends to underestimate the magnitude of the association of item and score more than *Rit* does when the item scale has four categories or more.

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

To cite this article: Metsämuuronen, J. (2020). 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

Introduction

In traditional educational and psychometrical test settings, we are interested in the reliability of the score or measurement scale, that is, the efficiency of the test score to discriminate between the lower- and higher-scoring test takers. The magnitude of reliability is strictly dependent on the discrimination power of single items in the compilation, that is, of the efficiency of the single items to discriminate between lower- and higher-scoring test takers (see Lord & Novick, 1968). The less the test items can discriminate the test takers from each other, the more items we need to reach high reliability and, in a parallel manner, the higher the discrimination power of the items, the shorter the test we can construct to reach the same efficiency. Therefore, we are interested in item discrimination power (IDP), the estimators of IDP, and the estimates they produce.

Operational definition of IDP

It is difficult to find a unanimous definition of IDP in the literature. The "definitions" tend to be as loose as that given by Lord and Novick (1968; see ETS, 2019; Liu, 2008; MacDonald & Paunonen, 2002) condensed above as "efficiency to differentiate between the lower and higher performing test takers". In order to define the concept in a manner that makes it possible to assess the possible under- and overestimation produced by different estimators of IDP, an important concept related to IDP, *deterministic discrimination*, is discussed here. Deterministic discrimination refers to an ultimate pattern where the score explains perfectly the behavior in the item, and then we expect to see the perfect explaining power (EP) between two variables ($\rho^2_{XY}=1$) that implies the perfect association ($\rho_{XY}=1$).

From the viewpoint of both the Pearson's product–moment correlation between the item *g* and the score *X* and Somers' *D* mainly discussed in this article, the perfect EP is achieved when the order of the cases both in the item and the score

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

are identical. Hence, an operational definition of the ultimate IDP could be as follows: *A test item is ultimately reliable when the score can predict the behavior of the test-takers in the test item in a deterministic manner*. A parallel practical definition of the ultimate IDP could be as follows: *The discrimination power of the test item is ultimate when, after arranging the test takers by the score or measurement scale, the item can discriminate the lower performing test takers from the higher performing test-takers in a deterministic manner*. If we observe a deterministic pattern, the index of IDP should be able to detect it. If, under these conditions, the estimators based on the correlations (such as biserial-, point-biserial-, polyserial-, or polychoric correlation) or ratios (such as Goodman–Kruskal Tau and Lambda, Somers' *D*, Pearson' Eta, or Kelley's discrimination index) give a value lower than 1, this indicates obvious underestimation of IDP. Similarly, the values higher than 1 indicate obvious overestimation of IDP. We may note that the indicators related to item response theory (IRT) and Rasch modelling, such as a-parameter (Birnbaum, 1968) and discrimination index by Verhelst, Glas, and Verstralen (1995), do not follow this logic. Basically, these indicators cannot reach the ultimate value because this condition presumes a dataset without stochastic error.

Some classical estimators reflecting the "true" IDP

In general, the indices of item discrimination summarize an item's relationship with a trait of interest (Moses, 2017). Within the classical test theory, we have several estimators for IDP; Oosterhof (1976) compared 19 of those (see also Cureton, 1966a, 1966b; ETS, 1960; Liu, 2008; Wolf, 1967). Two estimators based on the mechanics of Pearson's product–moment correlation coefficient (Pearson, 1896), the point-biserial coefficient of correlation also known as item–total correlation (ρ_{gX}, *Rit*), and item–rest correlation also known as the corrected item–total correlation coefficient (ρ_{gP}, *Rir*, Henrysson, 1963) may be the most frequently used indices of IDP. This can be deduced from the fact that they are set as default for the classical item analysis in the widely used general software packages such as SPSS, STATA, and SAS (see IBM, 2017; Stata corp., 2018; Yi-Hsin & Li, 2015). One challenge with these estimators is that, in real-life test settings, *Rit* or *Rir* cannot reach the value $\rho_{gX} = \rho_{gX} = 1$ because of the mismatch of the marginal scales of the item and the score. Underestimation may be drastic depending on the difficulty level of the item (see Metsämuuronen, 2016; 2017a, see also Table 1). From the deterministic pattern viewpoint it is pathological that, irrespective of the fact that the item would discriminate deterministically between the lower- and higher-scoring test takers and we would expect to see the perfect EP and the perfect IDP, the estimates by *Rit* and *Rir* approximate zero with the items of extreme difficulty level.

Another kind of challenge specifically with *Rit* is that the estimates are inflated because the score also includes the item of interest. Inflation has been characterized as "spurious" (e.g., Cureton, 1966b, p. 93; Howard & Forehand, 1962, p. 731; Wolf, 1967, p. 21). It is obvious that the less items we have comprising the score, the more effect has a single item in the score. Hence, in theory, using *Rir* or some other index correcting the inflation would be appropriate or even strongly recommended (see Liu, 2008). However, knowing that *Rit* may underestimate EP between the item and the score as much as 0.25 units of correlation or more (see Metsämuuronen, 2016, see Table 1)—and *Rir* underestimates IDP even more (see Metsämuuronen, 2017a, see Table 1)—inflation of a magnitude of 0.04–0.07 units of correlation may be taken as a secondary theoretical challenge.

The family of biserial (ρ_{BS}), polyserial (ρ_{PS}), and polychoric (ρ_{PC}) correlation coefficients (Pearson, 1900; 1913) assume that continuous and normally-distributed traits underlie dichotomized or polytomized items and polytomized scores. Crocker and Algina (1986) specifically suggest using ρ_{BS} instead of ρ_{gX} with the items of extreme difficulty level for overcoming the underestimation of ρ_{gX}. The magnitudes of the estimates by ρ_{BS}, ρ_{PS}, and ρ_{PC} tend to be higher than those by ρ_{gX}. However, of these, ρ_{BS} and ρ_{PS} tend to give obvious overestimation in certain cases. This is strictly seen in the formula of ρ_{BS} generalized from Olson, Drasgow, and Dorans (1982):

$$\rho_{BS} = \frac{\rho_{gX} \times \sigma_g}{\Phi(t)} \qquad (1)$$

where $\Phi(t)$ is the standard normal density related to the proportion of correct answers (*p*) and $\sigma_g = \sqrt{p(1-p)}$. The magnitude of the estimates by ρ_{BS} is maximal when $\rho_{gX} = 1$ and item variance is maximal, that is, when $\sigma_g = 0.5$ implying *p* = 0.5. Then, $t = \Phi^{-1}(0.5) = 0$, $\Phi(0) = 0.399$, and $\rho_{BS} = 1 \times 0.5/0.399 = 1.253$. From Eq. (1) we strictly infer that when the item variance is maximal, ρ_{BS} gives an obvious overestimate when $\rho_{gX} > 0.798$ (=0.399/0.5). On the other hand, ρ_{BS} and ρ_{PS} tend to underestimate item discrimination, especially with the items of extreme difficulty level. Seen in Eq. (1), the magnitude of ρ_{BS} and ρ_{PS} is strictly dependent of the magnitude of ρ_{gX}. Since the magnitude of the estimates by ρ_{gX} approximates zero with the extremely easy and difficult items, the magnitude of the estimates

by ρ_{BS} and ρ_{PS} also approximates zero irrespective of the fact that the score would explain the behavior in the item in a deterministic manner. To overcome the challenge of the obvious overestimation in ρ_{BS} and ρ_{PS}, Lewis, Thayer, and Livinstone (2003 cited in Livinstone & Dorans, 2004) developed a coefficient called *r*-polyreg correlation, an *r*-polyserial estimated by regression correlation. This coefficient 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). Other solutions have also been offered, for example, by Brogden (1949) and Henrysson (1971).

As with ρ_{BS} and ρ_{PS}, the magnitude of the estimates by ρ_{PC} tends to be higher than by ρ_{gX}. Additionally, the estimates by ρ_{PC} seem to produce more accurate reproduction of the measurement models used to generate the data (Holdago–Tello et al., 2010) and the estimates tend to produce unbiased estimated standard errors in the SEM analysis (Rigdon & Ferguson, 1991) even in small sample sizes (Flora & Curran, 2004). ρ_{PC} has been found to have advantages over the product–moment correlation coefficient in factor analysis (e.g., Holgado–Tello, Chacón–Moscoso, Barbero–García, & Vila-Abad, 2010) and specifically, in SEM analysis with ordinal datasets (e.g., Flora & Curran, 2004; Jöreskog, 1994; Rigdon & Ferguson, 1991; Uebersax, 2015) as well as in IRT and Rasch modeling (e.g., Forero & Maydeu-Olivares, 2009; Moustaki, Jöreskog, & Mavridis, 2004; Uebersax, 2015). However, a drastic challenge within the item analysis settings is that, by using ρ_{PC} instead of ρ_{gX}, we do not know what kind of composite the item discrimination refers to; we no longer refer to a known observed composite but, instead, a hypothetical composite the research is not privy to. A computational challenge is that, by using the established routines for estimating ρ_{PC} (e.g., Lancaster & Hamdan, 1964; Olsson, 1979; Tallis, 1962), the estimates cannot reach the extreme values +1 and –1.

All in all, the indices discussed above cannot identify deterministic (or near-deterministic) relationships between item score and total score. Some estimators provide us obvious underestimations and overestimations, and some refer to an unreachable test score.

Some directional estimators of IDP as an option for Rit and Rir

One challenge of the above-mentioned classical estimators, in addition, is that they do not take into account the assumed directional relationship (i.e., the assumption as noted, for example, in Byrne, 2001; Metsämuuronen, 2017b) between the item and the score: in psychometric theory, the overall trait being measured generally drives examinees' responses to, and, thus, scores on individual items.

There are directional measures of correlation that are consistent with the assumption that an overall trait drives examinee responses to individual items. Some of such measures are the directional coefficients Goodman–Kruskal Lambda and Tau (Goodman & Kruskal, 1954), Pearson's Eta coefficient (η) (Pearson, 1903, 1905), and Somers' *D* (Somers, 1962). These measures seem promising in overcoming the issues with the measures previously introduced though, as being nonparametric measures with a tendency to be less efficient than their parametric counterparts (see Metsämuuronen, 2017b; Öllerel & Croux, 2010; Siegel & Castellan, 1988), they tend to underestimate association between two variables. However, an advantage of the nonparametric measures, especially of Somers' *D*, is that it *can* reach the ultimate values correctly (Newson, 2002). Unlike the other directional indices mentioned above, Somers' *D* can identify both positive and negative monotonic deterministic relationships between item score and total score.

Behavior of the estimators: A practical example

Table 1 illustrates the behavior of the previous estimators of IDP by using two sets of items: a binary set (items A1, A2, A3) and a polytomous set (items B1, B2, B3). In both sets, one item follows deterministic pattern without stochastic error (A1 and B1) and we expect to see perfect item discrimination, while other items include stochastic error either to a minor (A2 and B2) or greater extent (A3 and B3). In all cases, the score discriminates the test takers in a deterministic manner.

From the data in Table 1 we see that both item-total and item-rest correlation underestimate IDP remarkably; this is seen clearly in the deterministic patterns (A1 and B1) where we expect to see $\rho_{gX} = \rho_{gP} = 1$ but observe values 0.53–0.75. Another character of the item-total correlation is also worth noting here. In the binary case, point-biserial correlation is factually a directional coefficient: the magnitudes of the estimates by η (score dependent) and ρ_{gX} are identical. Notably, ρ_{BS} exceeds the limits with item A1 (1.023). Though ρ_{PC} is not defined in the deterministic patterns (A1 and B1), its values seem to be close of those by Somers' *D* when the item discrimination is low (0.199 vs. 0.200 in A3 and 0.399 vs. 0.400 in B3). The value of Somers' *D* seems to be higher than ρ_{PC} with binary item (0.947 vs. 0.921 in A2) while the values of ρ_{PC} seems to be higher than Somers' *D* in the polytomous case (0.965 vs. 0.914 in B2) if item discrimination was high. Further research on the relation of ρ_{PC} and Somers' *D* may enrich our understanding.

Table 1. Example of the behaviour of selected directional coefficient of correlation

Test taker ID	A1	A2	A3	B1	B2	B3	(other items)	Score (X)
1	0	0	0	0	0	0	.	1
2	0	0	0	0	0	0	.	2
3	0	0	0	0	0	0	.	3
4	0	0	0	0	0	0	.	4
5	0	0	0	0	0	0	.	5
6	0	0	1	0	0	0	.	6
7	0	0	0	0	0	0	.	7
8	0	0	0	0	0	0	.	8
9	0	0	1	0	0	1	.	9
10	0	0	0	0	0	0	.	10
11	0	0	0	0	0	0	.	11
12	0	0	1	0	0	2	.	12
13	0	0	0	0	0	0	.	13
14	0	1	0	0	1	0	.	14
15	0	0	1	0	0	4	.	15
16	1	0	0	0	0	0	.	16
17	1	1	0	1	0	0	.	17
18	1	1	1	2	2	3	.	18
19	1	1	0	3	3	0	.	19
20	1	1	0	4	4	0	.	20
Item–total correlation (ρ_{qX})	0.751	0.711	0.150	0.659	0.636	0.326		
Item–rest correlation (ρ_{qP})	0.715	0.671	0.076	0.526	0.497	0.138		
Biserial and polyserial correlation (ρ_{RC})	1.023	0.959	0.205	0.931	0.898	0.460		
Polychoric correlation (ρ_{PC})	not dfnd	0.921	0.199	not dfnd	0.965	0.399		
Goodman & Kruskal Lambda (item dependent)	1	1	1	1	1	1		
Goodman & Kruskal Lambda (score dependent)	0.053	0.053	0.053	0.211	0.211	0.211		
Goodman & Kruskal Tau (item dependent)	1	1	1	1	1	1		
Goodman & Kruskal Tau (score dependent)	0.053	0.053	0.053	0.211	0.211	0.211		
Somers D (item dependent)	0.395	0.374	0.079	0.368	0.337	0.147		
Somers D (score dependent)	1	0.947	0.200	1	0.914	0.400		
Pearson Eta (item dependent)	1	1	1	1	1	1		
Pearson Eta (score dependent)	0.751	0.711	0.150	0.699	0.653	0.368		

A comparison of the directional coefficients in Table 1 shows that only Somers' *D* detects the deterministic pattern in the items. Though it seems that *L*, *T*, and η also can detect the deterministic item discrimination, it is just apparent. Namely, the values *L* (item dependent) = *T* (item dependent) = η (item dependent) = 1 refer to the behavior of the *score* instead of the item. We can infer this from items A2, A3, B2, and B3 with stochastic error: the stochastic error should cause a change in the value of the indices. However, among the directional measures, this change can be seen only in Somers' *D* (score dependent) and η (score dependent). Of these, η (score dependent) cannot detect the deterministic pattern in items A1 and B1, though the value is higher than *Rit* with the polytomous item.

All in all, Somers' *D* so directed that (paradoxically) the "score is dependent", that is, $D(g|X)$†, finds the ultimate IDP correctly and is capable of reaching both the negative and positive limits and, hence, has the potential of being considered as a serious alternative for *Rit* and *Rir*. We note the illogical wording of the direction with Somers' *D* in Table 1 in comparison with the traditional way of using the term "dependent." When, in the standard general linear modeling, we use the concepts of "dependent" and "independent," the "independent" factor (gender, for example) is used to explain the differences in the "dependent" variable (such as the score). Here, when the score is "dependent," we would expect that the item explains the differences in the score. However, it seems that this logic does not hold when using Somers' *D*: when the score is "dependent," the score explains the behavior in the items as inferred from Table 1.

Somers' D in the practical settings of educational measurement

Because Somers' *D* may be somewhat unfamiliar to some users in the practical educational measurement settings, relevant computational matters are discussed here. In general, Somers' *D* is used as a directional measure of association of two ordinal or continuous variables. We may be interested in knowing, as an example, how well the educational level explains the attitudes toward education or, vice versa, how well the attitudes explain the educational level (see the manual calculation in Metsämuuronen, 2017b). In the educational measurement settings, though, we maybe more interested in a specific direction of the association, that is, how well the score, i.e. the latent trait, explains the behaviour in an item rather than other way around.

As many nonparametric coefficients of association such as Kendall's tau and Goodman–Kruskall Gamma, Tau and Lambda, also Somers' *D* uses the concepts of concordance and discordance between the item (g) and the score (X) in the calculation. For the calculation, the concepts related to a *R* x *C* crosstable are usually used (see Siegel and Castellan, 1988; Metsämuuronen, 2017b); let us denote the dimension *R* for items and *C* for the score:

		Scale of the score				
Scale of the item		A_1	A_2	…	A_C	Sums of Rows
	B_1	n_{11}	n_{12}	…	n_{1C}	R_1
	B_2	n_{21}	n_{22}	…	n_{2C}	R_2
	…	…	…	…	…	…
	B_R	n_{R1}	n_{R2}	…	n_{RC}	R_R
	Sums of Columns	C_1	C_2	…	C_C	N

In the practical educational testing settings, if some of the general statistical software packages is in use, Somers' *D* is simple to calculate. In IBM SPSS, we select Analyze > Descriptive Statistics > Crosstabs... > Statistics… > Somers' D. The corresponding syntax in IBM SPSS is CROSSTABS /TABLES=item BY Score /STATISTICS=D. In SAS, the command PROC FREQ provides exact tests for Somers' *D* by specifying the SMDCR and SMDCR options in the EXACT statement. Correspondingly, RStudio, as an example, uses the syntax SomersDelta (x, y = NULL, direction = c("row", "column"), conf.level = NA, ...).

However, if no general package of statistical tools is in use, the manual calculation starts by ordering the test takers by the score *X*. Assume that we are interested in the specific binary item *g*—naturally we are interested all of the items and these are not restricted to binary ones though here we use this as an example. Assume that we would have obtained the following dataset with *N* = 10 test takers ordered by the score—Johan scores the highest (21 points) and John the lowest (0 points).

Test taker	John	Jill	James	Jonah	Jennifer	Jenny	Jacob	Judy	Jane	Johan
Score *X*	0	1	4	4	10	14	15	16	20	21
Item *g*	0	1	0	1	0	0	1	0	1	0

† Note here the untraditional direction of condition ($g|X$) usually used with Somers' *D*. This notation corresponds with the traditional thinking of condition: "*g* in condition of *X*", that is, *g* is dependent on *X,* that is *"g* dependent*"*. However, with Somers' *D,* this notation is called "*X* dependent". In this article, the specific notation $D(g|X)$ refers to "*g* dependent" but, in the formulae, the traditional notation of Somers' *D* is used. This unorthodox marking is noted when a possibility for an imminent misunderstanding occurs.

We form a frequency table of the score and the item:

		Score X									
		0	1	4	10	14	15	16	20	21	marg. distr. of *g*
item g	0	1	0	1	1	1	0	1	0	1	6
	1	0	1	1	0	0	1	0	1	0	4
	marg. distr. of *X*	1	1	2	1	1	1	1	1	1	10

For calculating Somers' *D*, all pairs of observations are compared and the sums of concordant pairs (*P*) and disconcordant pairs (*Q*) are formed. If a pair of observations g_i and g_j and corresponding X_i and X_j have ranks in the same direction, the pairs are concordant. We denote each cell frequencies by n_{ij}. For the concordant pairs, we calculate how many observations there are in the cells below and to the *right* of the cell n_{ij}. These are denoted by N_{ij}^{+}. For the first cell n_{11}, we get $N_{11}^{+} = 1+1+0+0+1+0+1+0 = 4$. For the next cell the value is $N_{12}^{+} = 1+0+0+1+0+1+0 = 3$ and so on until the last one $N_{18}^{+} = 0$. Correspondingly, the disconcordant pairs denoted by N_{ij}^{-} are found in the cells below and to the *left* of the cell n_{ij}. The first ones are related to the cell n_{12} and n_{13}: $N_{12}^{-} = 0$ and $N_{13}^{-} = 0+1 = 1$. All possible values for N_{ij}^{+} and N_{ij}^{-} are computed and these are multiplied by the related n_{ij}. In the example, the number of all the pairs in the same direction, that is, the concordant pairs, is

$n_{11}N_{11}^{+} = 1\times(1+1+0+0+1+0+1+0) = 4$, $n_{12}N_{12}^{+} = 0\times(1+0+0+1+0+1+0) = 0$, $n_{13}N_{13}^{+} = 1\times(0+0+1+0+1+0) = 2$,
$n_{14}N_{14}^{+} = 1\times(0+1+0+1+0) = 2$, $n_{15}N_{15}^{+} = 1\times(1+0+1+0) = 2$, $n_{16}N_{16}^{+} = 0\times(0+1+0) = 0$, $n_{17}N_{17}^{+} = 1\times(1+0) = 1$,
$n_{18}N_{18}^{+} = 0\times(0) = 0$, altogether $P = \sum_{ij} n_{ij}N_{ij}^{+} = 4+0+2+2+2+0+1+0 = 11$.

Correspondingly, the number of all the pairs in the opposite order is:

$n_{12}N_{12}^{-} = 0\times(0) = 0$, $n_{13}N_{13}^{-} = 1\times(0+1) = 1$, $n_{14}N_{14}^{-} = 1\times(0+1+1) = 2$, $n_{15}N_{15}^{-} = 1\times(0+1+1+0) = 2$,
$n_{16}N_{16}^{-} = 0\times(0+1+1+0+0) = 0$, $n_{17}N_{17}^{-} = 1\times(0+1+1+0+0+1) = 3$, $n_{18}N_{18}^{-} = 0\times(0+1+1+0+0+1+0) = 0$,
$n_{19}N_{19}^{-} = 1\times(0+1+1+0+0+1+0+1) = 4$, altogether $Q = \sum_{ij} n_{ij}N_{ij}^{-} = 0+1+2+2+0+3+4 = 12$.

By using the concepts of *P* and *Q*, Somers' *D* (item dependent) can be calculated as

$$D_{R|C} = D_{\text{item dependent}} = \frac{2(P-Q)}{N^2 - \sum_{j=1}^{c}\left(n_{Cj}{}^2\right)} \tag{2}$$

where n_{Ci} is the number of cases in the categories $c = j$ related to score *X*. In the case,

$$D_{R|C} = \frac{2(P-Q)}{N^2 - \sum_{j=1}^{c}\left(n_{Cj}^2\right)} = \frac{2(11-12)}{10^2 - \left(1^2+1^2+2^2+1^2+1^2+1^2+1^2+1^2\right)} = \frac{-2}{89} = -0.022 \tag{3}$$

Parallel, Somers' *D* (score dependent) is

$$D_{C|R} = D_{\text{score dependent}} = \frac{2(P-Q)}{N^2 - \sum_{i=1}^{r}\left(n_{Ri}{}^2\right)} = \frac{-2}{10^2 - \left(6^2+4^2\right)} = \frac{-2}{48} = -0.042 \tag{4}$$

Of these, the latter tells (though paradoxically named) the behavior of the item and the former indicates the behavior of the score (see Table 1). Either way, the item *g* of interest cannot discriminate the test takers from each other which was obvious from the dataset.

The magnitude of Somers' *D* varies within −1 and +1 and, hence, unlike Goodman–Kruskal Gamma, Lambda and Tau, it can reach the negative values as well as positive values. With regard to item analysis, *D* = 1 always when the order of the test takers in the item is the same as that in the score regardless of the number of cases, dimensions in the item and the score, the number of tied values, difficulty levels in the items, or the number of items on the test. In such cases, all the pairs of g_i and g_j and corresponding X_i and X_j are in the same direction, and therefore, the maximum value of the directional rank correlation is obtained. The value *D* = −1 is obtained when the order of the test takers in the item is

opposite in comparison with the score. The value $D = 0$ is obtained when the number of concordant pairs equals the disconcordant pairs as in the example above.

We may recall that Somers' D has a long history in item analysis though it is not much discussed. Namely, the robust rank–biserial coefficient of correlation ρ_{RB} (Cureton, 1956; see also Glass, 1966, who derived the same coefficient from different grounds) related to the nonparametric and directional U-test statistic (Wendt, 1972) is shown to be a special case of Somers' D (Newson, 2008). While ρ_{RB} is restricted to binary case, Somers' D can be used also with the polytomous items. Because of the connection with U-test statistics, we know that Somers' D directed so that the score explains the behavior in the item, that is, Somers' $D(g|X)$, essentially, indicates the slightly modified proportion of correctly ordered test takers in the item after they are ordered by the score. This fits quite well with operational definitions of IDP above.

Research question

From the practical example above and the general behavior of Somers' D it is known that this specific coefficient could be a potential alternative for item–total and item–rest correlation in reflecting IDP. The dataset in Table 1 is small and a theoretical one. We do not know how well Somers' D behaves in the real-life datasets with varied test difficulty, reliability, number of items, number of test takers, and the marginal distribution of the score as well as with items with varying difficulty levels, discrimination power, and the marginal distribution of the score. These matters are studied in detail through a simulation with empirical datasets.

Methodology

The simulation dataset

The characters of Somers' $D(g|X)$—hence forth D—in comparison with *Rit* —hence forth R—are studied with a simulation of 13,392 items from 1,296 tests with varying characteristics based on different combinations of randomly selected test takers from an unpublished national-level dataset of 4,500 grade 9 test takers of a mathematical proficiency test (FINEEC, 2018). The original set of the data was used to prepare several smaller datasets with varied difficulty levels ($\overline{p}$), magnitude of reliability (α), test lengths (k), number of cases (N), and degrees of freedom in the item, $df(g)$ = number of marginal categories minus 1 and in the score, $df(X)$ = number of marginal categories minus 1.

The original real-world dataset did not include extremely difficult items. Hence, two kinds of datasets were constructed for the simulation—those based on real-world test takers (83%) and those based on artificial ones (17%). In the first phase, three sets of 10 random samples of sizes n = 200, 100, and 50 with 30 dichotomous items were picked from the original real-life dataset. Two additional artificial 30-item datasets—an extremely difficult one and a moderately difficult one—were created to enrich the data. Hence, after the first phase, the number of datasets totaled 3 × 12 = 36.

In the second phase, eight shorter tests were constructed based on the 36 datasets with 30 items by varying the degrees of freedom in the score. These tests comprised 20, 21, 22, 24, 26, 27, 28, and 30 items. After the second phase, the total number of tests was 36 × 8 = 288 partly dependent tests. In the third phase, each of the 288 tests was used to create tests with the varied degrees of freedoms of the items. The test with 21 binary items is used as an example of the logic.

A set of 21 binary items with gradually increasing difficulty levels can be divided into three sets of subtests related to the same score. One of these is the traditional test including all 21 items as separate binary items, that is, a test of 21 binary items with the range of 0–1 in the item scale (k = 21; $df(g)$ = 1). The other extreme is the test with three "parallel" sums of every third item. These three scores formed three "items" with the range of 0–7 in the item scale (k = 3; $df(g)$ = 7). The third subtest comprised one test of seven parallel sums of every seventh item with the range of 0–3 in the scale (k = 7; $df(g)$ = 3). Hence, with the original 21 items, three tests were formed, and these produced 31 items with different characteristics: 21 items with $df(g)$ = 1, seven items with $df(g)$ = 3, and three items with $df(g)$ = 7.

Similarly, the sets of 22, 26, and 27 items produced three such "parallel" sets of tests, while those of 20, 24, 28, and 30 items produced 5, 7, 5, and 7 sets of tests, respectively—altogether 36 tests in each 3 × 12 dataset. Thus, a total of 36 × 36 = 1,296 tests was produced with varied values of test difficulty, reliability, number of items, number of test takers, and $df(X)$ (see Table 2). Consequently, the procedure provided us with 13,392 items with varying difficulty levels, discrimination power, and $df(g)$ (see Table 3).

Table 2. Selected characteristics *of 1,296 tests in simulation*

Difficulty level (average *p*)	**Description**	**Average reliability (α)**	**Average *R***	**Average *D***	**Number of datasets**	**Remark**
0 – 0.299	Extremely difficult	0.901	0.785	0.766	47	Artificial
0.3 – 0.399	Very difficult	0.927	0.809	0.770	112	Artificial/Real-world
0.4 – 0.499	Difficult	0.956	0.863	0.826	57	Real-world
0.5 – 0.599	Mediocre	0.833	0.720	0.652	142	Real-world
0.6 – 0.699	Easy	0.867	0.728	0.666	721	Real-world
0.7 – 0.799	Very easy	0.863	0.731	0.673	217	Real-world
Total		0.873	0.743	0.685	1,296	

Table 3. Selected characteristics of 13,392 items in simulation

***df*(*g*)**	**Number of items**	**Average *R***	**Average *D***	**Average item difficulty ($\bar{p}$)**	**Average item variance ($\bar{\sigma}_g^2$)**
1	7131	0.5063	0.6284	0.6097	0.2108
2	2715	0.6463	0.6698	0.6125	0.5149
3	1233	0.7266	0.7035	0.6169	0.9190
4	658	0.7876	0.7369	0.6147	1.4418
5	415	0.8230	0.7535	0.6009	2.1025
6	335	0.8569	0.7779	0.6224	2.9173
7	234	0.8832	0.7996	0.6065	3.8762
8	123	0.9032	0.8150	0.6320	5.1814
9	165	0.9197	0.8363	0.5933	6.6778
10	140	0.9319	0.8479	0.5894	7.9771
11	93	0.9427	0.8606	0.6319	9.5838
12	74	0.9494	0.8670	0.6112	11.5998
13–15	76	0.9488	0.8637	0.6231	12.7240
Total	13,392	0.8479	0.7918	0.6156	6.0680

Data analysis

The dataset is analyzed by using basic tools to compare *D* and *R*. The main illustrative tool is the difference between *D* and *R*. When $D - R > 0$, obviously, the magnitude of the estimate by *D* is higher than that by *R*. While knowing that *R* always underestimates IDP in the real-life testing settings, the values of $D - R < 0$ are indicative that *D* underestimates IDP even more than *R*. We remember that the magnitude of estimates by item-rest correlation coefficient (*Rir*) is lower than that by R and, hence, *Rir* underestimates IDP more than *R*. Therefore, the condition $D - R > 0$ is indicative that the magnitude of the estimates is in the order $D > R > Rir$.

Findings

D vs. R with binary items

The first thing to note before the simulation is that, with deterministic patterns, irrespective of *df*(*g*) and *df*(*X*), *D* gives exact and correct estimate of IDP while *R* underestimates IDP practically always. An extreme estimate $D = 1$ will always be obtained when the order of cases in the score is identical with the order in the item as discussed above. As noted above, the perfect $R = 1$ can be reached only when *df*(*X*) = *df*(*g*), which is a highly specific condition.

In the real-life settings, the estimates by *D* tend to underestimate IDP less than *R* does in the binary datasets as well as when *df*(*g*) = 2 (Table 3 and Figures 1 and 2; see also Figure 4). With binary items ($n = 7\,131$), in 99.9% of the estimates the magnitudes of the estimates by *D* are higher than those by *R* (Figure 1). Similarly, with items with three categories in the marginal distribution ($n = 2{,}715$), in 81.2% of the estimates the magnitude of the estimate by *D* is higher than the estimate by *R* (Figure 2). Specifically, when items are extremely easy or extremely difficult, where the underestimation in *R* is the highest, *D* gives a notable advantage over *R* in reflecting the IDP of the item. This does not indicate that the value of *D* would be true or correct, or that IDP will be reflected with the same intensity as it happens in cases with ultimate discrimination $D = 1$. However, because it is difficult to think how *D* could overestimate IDP due to its being a nonparametric coefficient, the higher values in *D* when *df*(*g*) = 1 and 2 indicate that the estimates of IDP by *D* are *closer to the true value* in comparison with the estimates by *R*.

From Figure 1, we note an apparent asymmetricity in the U-shaped pattern of *D* – *R*. The reason is the too perfect artificial datasets. The right-hand side extreme leading to discrepancy of around 0.60 units of correlation is based on the real-life datasets. In the other extreme in the left-hand side with around 0.40 units of correlation the lower values are based on artificially high item-total correlations.

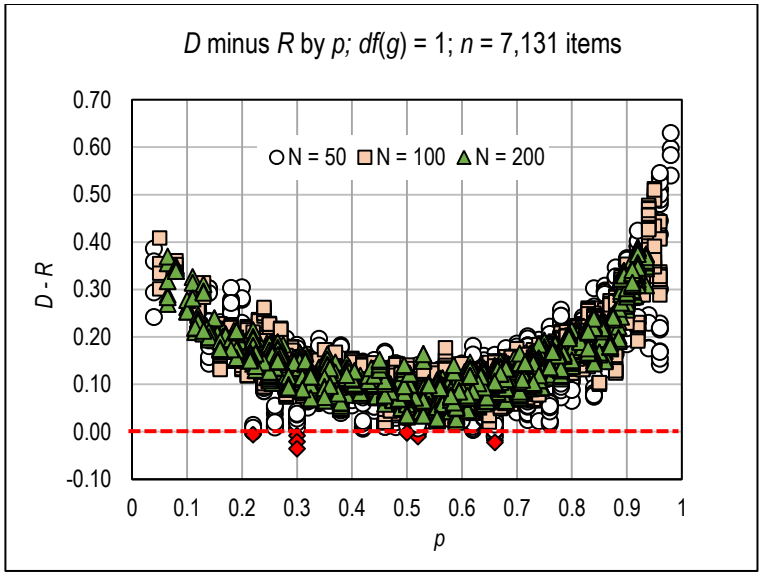

Fig. 1. Difference between the estimates by D and R; df(g) = 1

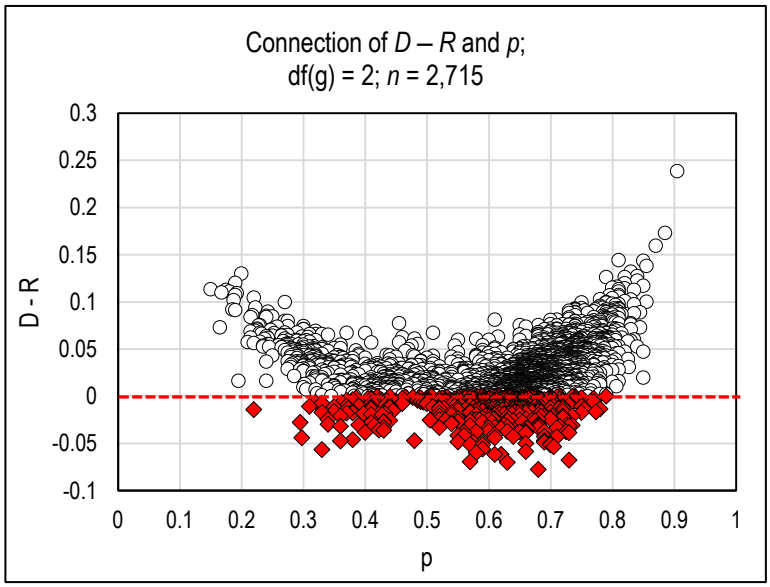

Fig. 2. Difference between the estimates by D and R; df(g) = 2

D vs. R with polytomous items

The simulation with the real-life items shows that when *df*(*g*) > 2 and when the number of the marginal categories in the item increases, *R* seems to be superior than *D* in reflecting IDP. Though the estimates by *D* are always exact in deterministic patterns of item discrimination irrespective of the number of categories in the items, in real-world datasets, the underestimation in *D* is evident (see Table 3 and Figures 3 and 4). With *df*(*g*) = 3, of the 1,233 estimates by *R*, 84% are higher in magnitude than those by *D.* With *df*(*g*) = 4, that is, in a typical Likert type scale, 98% of the estimates (*n* = 656) by *R* are higher in magnitude than those by *D*. It is evident that when *df*(*g*) > 2 and the number of categories in the item increases, *R* is superior over *D* reflecting the item discrimination.

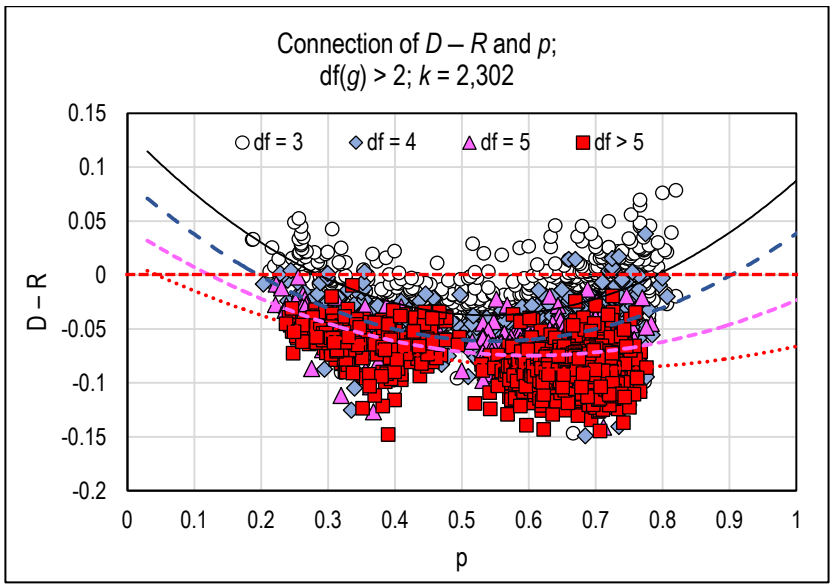

Fig. 3. Difference between the estimates by D and R; df(g) > 2

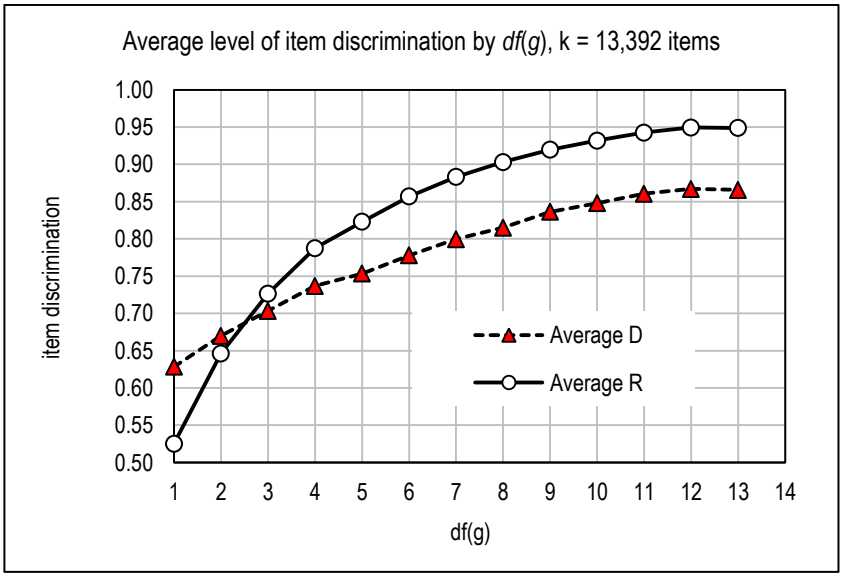

Fig. 4. Underestimation in D in relation with R as a function of df(g) based on Table 3

Apparently, seen in Figure 4, the limitations of the original dataset in the simulation may mislead us to assume that the underestimation in *D* increases systematically with *df*(*g*). However, this is not true. Even when the degrees of freedom of the item are indefinitely high—that is, when the item is a continuous one—both *D* and *R* can reach the maximal value 1. Behind this, we find the basic reality related to the mechanical connection of item and the score. In the measurement modelling settings, the association of *g* and *X* is, in fact, mechanically determined because the score is a compound of the items. As known, Somers' *D* can reach the ultimate magnitude *D* = 1 always when the order of the cases in the score and in the item are identical. Notably, also *R* approximates the ultimate magnitude *R* = 1 when the item scale approximates the scale of the score. This is understandable when we remember the relation between the items and the score. If there would be only one item on the test, the correlation between the item and the "score" formed by this item would be, obviously, perfect. With *two* items on the test with the wide scale, usually interpreted as "parallel tests", they both reflect (approximately) perfectly the (double-lengthen) total score. The more we have items on the test and the less there are categories in the items, the less the single item reflects the score.

Discussion

Advances of Somers' D

Newson (2002) notes three general advantages of Somers' $D(X|Y)$ over the Pearson product-moment correlation. First, Somers' $D(X|Y)$ has a favorable characteristic to reach the values +1, 0, and −1 accurately. Second, Somers' $D(X|Y)$ is more robust in extreme observations and nonlinearity. Third, the interpretation of Somers' $D(X|Y)$ is straightforward, and it may be easier to interpret in words—within item analysis, the value of $D(g|X)$ refers strictly to the (slightly modified) proportion of correctly located test takers in the item after they are ordered by the score. Additionally, the directional nature of Somers' $D(g|X)$ may be an advantage in the measurement modeling settings as discussed above. In the applied settings of measurement modeling, the specific form of Somers' $D(g|X)$, which corresponds to the logic used in item analysis, reflects this assumption better than *Rit* and *Rir*; with the latter two, the direction of the effect is not defined.

Combining the advantages Newson (2002) points out, the discussion in the introduction part of this article, and the empirical findings in this article, *D* could be proposed as a "superior alternative for *Rit*"—as well as *Rir* because the latter underestimates IDP even more than the former—in reflecting true IDP with dichotomous items because of the following reasons:

1. *D* reaches the values +1 and −1 accurately while *Rit* and *Rir* cannot reach the limits within practical measurement modeling settings.
2. *D* is more robust for the extreme observations and for nonlinearity than *Rit* and *Rir*.
3. *D* is superior to *Rit* and *Rir* with the dichotomous items, because when $df(g) = 1$, it is highly probable that *D* produces an estimate that underestimates IDP less than *Rit* and *Rir* does.
4. *D* has a logical directional nature from the modern measurement–modeling viewpoint; while *Rit* and *Rir* tell about the unspecified association of the variables, *D* tells us how well the latent factor (score) explains the behavior in the manifested variable (item).
5. *D* increases the possibilities of detecting the maximally discriminating test items in comparison with *Rit* and *Rir*; $D = 1$ when the order of the test takers in the item is the same as in the score 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. Of these cases, *Rit* and *Rir* can reach the ultimate value only when $df(X) = df(g)$.

When it comes to ρ_{PC},

6. *D* 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 in research.
7. *D* can reach the ultimate values +1 and −1 while, by using the standard procedures, ρ_{PC} cannot find solution in the deterministic patterns.
8. *D* 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 the cross-tabulation and normality of the phenomenon.
9. *D* is easy to calculate manually in practical test settings such as classroom testing, while calculation of ρ_{PC} requires specific software packages and complex procedures.

Hence, for several reasons, Somers' *D* could be a good alternative for *Rit* and *Rir* and to some extent also for ρ_{PC}. All in all, it would be safe—or may be even recommendable—to use *D* as an alternative to *Rit* and *Rir* as an estimator of IDP with dichotomous items. Though *D* reaches the ultimate values of IDP accurately, the estimates seem to include underestimation in certain conditions. It is obvious from the simulation that when the number of categories in the marginal distribution of the item exceeds 3, the probability of finding the estimates higher in magnitude by ρ_{gX} than by *D*. Underestimation by *D* is also hinted by Göktaş and İşçi (2011) in their simulation with the same scale in both variables ($df(Y) = df(X) = 3$). The lower values by *D* are understood because of Greiner's relation (Greiner, 1909), discussed by Kendall (1949) and Newson (2002). Greiner's relation shows that if the scales of both the item and the score are continuous, except the values ± 1 and 0, the magnitude of the estimates by Somers' *D* would be lower than those by Pearson correlation.

Limitations of the study

Though the numbers of subtests and items used in the simulation are rather convincing, three main limitations of the study are raised here related to the simulation datasets. First, the whole set of 1,296 datasets and 13,392 items in the simulation is based on one basic dataset related to one specific grade (9) in a specific country with high results in mathematics. The limitations of this dataset were not studied further because the original dataset of 4500 students was used just as a "population" or "factory" for the smaller tests. It is possible that the results may have been somewhat different if some other grade or different test items would be used in the simulation. Most probably, however, the number of items is high enough to find reasonable variation between *D* and *R*.

Second, because of the procedure of producing the smaller datasets, many of the items are somehow related to each other because they relate with the same score. We do not know, what is the real effect of this to the results. The magnitudes of the estimates within the "family" of the same test score do not differ from each other radically. Hence, the 13,392 estimates are not fully independent from each other, but they offer a kind of jittered dataset in some extent. Nevertheless, the idea in the study was to compare the behavior of the generally known indices in the real-life settings; now the relative magnitude of the estimates was studied in wide variety of different type of items regardless the possible relatedness of the tests and test items. Independent datasets would, most probably, confirm the same result. Replications of the design or another approach with more independent estimates may increase or knowledge of the matter.

Third, the quality of the artificial datasets is questionable. This part of the dataset seems to be too perfect in comparison with the real-world datasets. This is seen as asymmetric distribution of residuals in Figure 1 and in few cases of binary items where the magnitude of *R* was higher than the value of *D*; these all came from the artificial combination of extreme difficulty level and high item–total correlation. From this point of view, wider simulations with extremely difficult real-life datasets would benefit us.

Suggestions for the further studies

Three directions of continuing studied with IDP are proposed based on the results. First, this article focused on comparing *D* and point-biserial correlation from the viewpoint of their capability to reflect the real IDP. Parallel analysis with bi- and polyserial and polychoric correlation coefficients or with *r*-polyreg correlation would enrich our knowledge of the matter.

Second, the systematic nature of underestimation related to the higher degrees of freedom of item (see Figure 3 and the systematic decrease in the graph) hints that it could make sense to derive a "dimension-corrected Somers' *D*" that could be even more useful tool in estimating IDP. While *D* is (one of) the superior option(s) to use instead of *Rit* and *Rir* in dichotomous cases, the dimension-corrected *D* could be a good tool in polytomous cases, specifically for item-analysis purposes but may also be wider in scope.

Third, knowing that point-biserial correlation is embedded to all generally known estimators of reliability and that the estimates underestimate always the "real" IDP, it may motivate us to seek the "real" reliability by replacing ρ_{gX} with some other estimator of IDP that could be "superior" than *Rit* and *Rir*. Studies in this area may enrich the discussions and practices within measurement modeling settings.

Acknowledgements

The writer sincerely thanks Dr. Roger Newson, research associate at the Faculty of Medicine, School of Public Health, Imperial College, London, for suggesting Greiner's relation to understand the underestimation in Somers' *D*. He also helped in getting access to other useful resources on the topic. Also, sincere thanks to Assistant Professor R. Philip Chalmers from the York University, Toronto, for pointing to the challenges of the polyserial correlation coefficient in the measurement settings in a private discussion about handling the alternative ways of estimating reliability of the test score.

References

Birnbaum A (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397–479). Addison-Wesley Publishing Company.

Brogden, H. E. (1949). A new coefficient: Application to biserial correlation and to estimation of selective efficiency. *Psychometrika, 14*(3), 169–182. https://doi.org/10.1007/BF02289151

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

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* Wadsworth.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika, 16*(3), 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. (1966a). Simplified Formulas for Item Analysis. *Journal of Educational Measurement*, *3*(2), 187–189. https://doi.org/10.1111/j.1745-3984.1966.tb00879.x

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

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

ETS (2019). Glossary of Standardized Testing Terms. https://www.ets.org/understanding_testing/glossary/

FINEEC (2018). *National Assessment of Learning Outcomes in Mathematics at Grade 9 in 2004*. Unpublished dataset opened for the re-analysis 18.2.2018. Finnish National Education Evaluation Centre.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*(3), 275-299. https://doi.org/10.1037/a0015825

Glass, G. V. (1966). Note on rank biserial correlation. *Educational and Psychological Measurement*, *26*(3), 623-631. https://doi.org/10.1177/001316446602600307

Goktas, A. & Isci, O. A. (2011). Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation. *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). Über 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.

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 ed.) (pp. 130–159). American Council on Education.

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*, 153–166. https://doi.org/10.1007/s11135-008-9190-y

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. (2017). IBM SPSS Statistics 25 Algorithms. IBM. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/25.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf

Jöreskog, K. G. (1994). Structural equation modeling with ordinal variables. In T. W. Anderson, K. T. Fang, & I. Olkin (Eds.), *Multivariate analysis and its applications* (pp. 297–310)*.* Hayward, CA: Institute of Mathematical Statistics. https://doi.org/10.1214/lnms/1215463803

Kendall, M. (1949). Rank and Product–Moment Correlation. *Biometrika, 36*(1/2), 177–193. https://doi.org/10.2307/2332540

Lancaster, H. O., & Hamdan, M. A. (1964). Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters. *Psychometrika, 29*(4), 383–391. https://doi.org/10.1007/BF02289604

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

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

Metsämuuronen, J. (2016). Item–total Correlation as the Cause for the Underestimation of the Alpha Estimate for the Reliability of the Scale. *GJRA - Global Journal for Research Analysis*, *5*(1), 471–477. https://www.worldwidejournals.com/global-journal-for-research-analysis-GJRA/file.php?val=November_2016_1478701072__159.pdf.

Metsämuuronen, J. (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.

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

Moustaki, I, Jöreskog, K. G., & Mavridis D. (2004). Factor Models for Ordinal Variables with Covariate Effects on the Manifest and Latent Variables: A Comparison of LISREL and IRT Approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(4), 487–513. https://doi.org/10.1207/s15328007sem1104_1

Newson, R. (2002). Parameters Behind "Nonparametric" Statistics: Kendall's tau, Somers D and Median Differences. *The Stata Journal, 2*(1), 45–64. http://www.stata-journal.com/sjpdf.html?articlenum=st0007

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

Ollerer, V., & Croux, C. (2010). Robust high-dimensional matrix estimation. In K. Nordhausen & S, Taskinen (Eds.), *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja* (pp. 325–350). Springer.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika, 47*(3), 337–347. https://doi.org/10.1007/BF02294164

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*(4), 443–460. https://doi.org/10.1007/BF02296207

Oosterhof, A. C. (1976). Similarity of various item discrimination indices. *Journal of Educational Measurement*, *13*(2), 145–150. https://doi.org/10.1111/j.1745-3984.1976.tb00005.x.

Pearson, K. (1896). Mathematical contributions to the theory of evolution III. regression, heredity, and panmixia. *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. https://archive.org/details/ongeneraltheory00peargoog/page/n3.

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

Rigdon, E. E., & Ferguson, C. E. JR. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28*(4), 491–497. https://doi.org/10.1177/002224379102800412

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

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

Tallis, G. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics, 18*(3), 342–353. https://doi.org/10.2307/2527476

Uebersax, J. S. (2015). The tetrachoric and polychoric correlation coefficients. *Statistical Methods for Rater Agreement.* http://www.john-uebersax.com/stat/tetra.htm

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

Verhelst ND, Glas CAW, & Verstralen HHFM (1995). *One-parameter logistic model OPLM*. Cito.

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

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