Behaviormetrika (2021) 48:283–307
https://doi.org/10.1007/s41237-021-00138-8

ORIGINAL PAPER

Directional nature of Goodman–Kruskal gamma and some consequences: identity of Goodman–Kruskal gamma and Somers delta, and their connection to Jonckheere–Terpstra test statistic

Jari Metsämuuronen[1,2]

Received: 13 August 2020 / Accepted: 23 March 2021 / Published online: 10 April 2021

Abstract

Although usually taken as a symmetric measure, G is shown to be a directional coefficient of association. The direction in G is not related to rows or columns of the cross-table nor the identity of the variables to be a predictor or a criterion variable but, instead, to the number of categories in the scales. Under the conditions where there are no tied pairs in the dataset, G equals Somers' D so directed that the variable with a wider scale (X) explains the response pattern in the variable with a narrower scale (g), that is, $D(g \mid X)$. Hence, $G = G(g \mid X) = D(g \mid X)$ but $G \neq D(X \mid g)$ and $G \neq D$(symmetric). If there are tied pairs, the estimates by $G = G(g \mid X)$ are more liberal in comparison with those by $D(g \mid X)$. Algebraic relation of G and D with Jonckheere–Terpstra test statistic (JT) is derived. Because of the connection to JT, $G = G(g \mid X)$ and $D = D(g \mid X)$ indicate the proportion of logically ordered test-takers in the item after they are ordered by the score. It is strongly recommendable that gamma should not be used as a symmetric measure, and it should be used directionally only when willing to explain the behaviour of a variable with a narrower scale by the variable with a wider scale. This fits well with the measurement modelling settings.

Keywords Goodman–Kruskal gamma · Somers D · Jonckheere–Terpstra test · Kendall *tau-b* · Item analysis

Communicated by Kohei Adachi.

✉ Jari Metsämuuronen
jari.metsamuuronen@gmail.com

[1] Finnish Education Evaluation Centre, Mannerheiminaukio 1 A, 6th Floor, P.O. Box 28, 00101 Helsinki, Finland

[2] NLA University College, Bergen, Norway

1 Introduction

1.1 Family of gamma, delta, and tau

For estimating the association between two ordinal-scaled variables, two approaches are usually used: the one based on covariance and the one based on probability. The approach using covariance includes such widely used estimators as product–moment correlation coefficient (PMC; Pearson 1896 onwards) originally meant for two observed continuous variables and polychoric correlation (RPC; Pearson 1900, 1913) for two unobservable latent variables. Within the approach using probability, the most commonly used measures of association come from the family that includes Kendall's *tau-a* and *tau-b* (Kendall 1938, 1948), Goodman–Kruskal gamma (*G*; Goodman and Kruskal 1954), and Somers' delta (*D*; Somers 1962). Also, such rarely used estimators as Kim's *d*y.x (Kim 1971) and Wilson's *e* (1974) are part of this family. As a family of coefficients, it is usually referred to either as tau family (e.g., Kendall 1948; Kendall and Gibbons 1990), gamma family (e.g., Van der Ark and Van Aert 2015; Woods 2007), or delta family (e.g., Newson 2006; Metsämuuronen 2020a, b). Kendall's *tau-a* can be taken as the mother of the other estimators because, when there are no tied pairs between the variables, they all equal with *tau-a* (see Kendall and Gibbons 1990; Newson 2006). This article studies, specifically, the characteristics of *G* and shows that, under certain conditions, *G* is a special case of *D* although sometimes the opposite is suggested (e.g., Kvålseth 2017).

1.2 Some known characteristics of G within the measurement modelling settings

G and partial *G* (Goodman and Kruskal 1954; Davis 1967) are used, although rarely, it seems, in measurement modelling settings (see, e.g., Forthmann et al 2020; Kreiner and Christensen 2009; Nielsen and Santiago 2020). However, *G* has some favourable characteristics related to these settings. Namely, in comparison with the wider used estimator PMC, *G* appears to be robust against many sources of so called systematic mechanical error (SME) causing mechanical underestimation of association (Metsämuuronen 2021). SME is a new concept related to estimators of association referring to fact that the estimates of association include error that is mechanical in nature and it occurs in a systematic manner in certain estimators of association in varying quantity. For example, while PMC is notably affected by such sources of SME as restriction of range in general (see the literature in, e.g., Mendoza and Mumford 1987; Sackett and Yang 2000; Sackett et al 2007 and simulations by Martin 1973, 1978; Olsson 1980), item difficulty, the number of categories in the item and in the score, and the distributions of the latent variables, *G* produces estimates that are SME-free in all of these conditions (see simulations in Metsämuuronen 2021). In practical terms, while PMC always underestimates the true association because of mechanical reasons, *G* reflects the true association without loss of information caused by the mechanical reasons regardless of the sources of SME mentioned above. Hence, *G* appears to be a surprisingly interesting coefficient in resisting SME in the estimation of association. However, although *G* is accurate in reflecting the true association between two variables, it has two opposite challenges: obvious underestimation when the number of categories in the variables gets high and possible inflation magnitude of the estimates. These are discussed below.

Because being based on probability, the embedded linear nature in *G* in comparison with the estimators with trigonometric nature such as PMC makes *G* underestimate the association between an item and the score in an obvious manner (see Metsämuuronen 2021). The phenomenon is similar with Somers' *D* (Metsämuuronen 2020b; Göktaş and İşçi 2011), and it can be explained by Greiner's relation (Greiner 1909) discussed by Kendall (1948), Newson (2002), and Metsämuuronen (2020b, 2021). Greiner's relation states that, with continuous variables *X* and *Y*, $tau\text{-}a = G = D$ and, then, PMC between variables *X* and *Y* equals $\rho_{XY} = \sin\left(\frac{1}{2}\pi \times tau_a\right)$ $= \sin\left(\frac{1}{2}\pi \times G\right) = \sin\left(\frac{1}{2}\pi \times D\right)$. Consequently, with continuous variables, the values by ρ_{XY} of 0, $\pm 1\big/\sqrt{2}$, and ± 1 as examples correspond with the values by *G* and *D* of 0, $\pm 1/2$, and ± 1, respectively. Then, except for the extreme values ± 1 and 0, the magnitude of the estimates by ρ_{XY} tends to be greater than those by $G = D$. While it is known that *D* underestimates association of an item and the score when the number of categories in the item exceeds three (Metsämuuronen 2020b), *G* seems to underestimate association when the number of categories exceeds four (Metsämuuronen 2021).

Another discussed challenge in *G* is its possible inflation in the estimates. Kvålseth (2017) notes that the estimates by *G* "may be highly inflated making it incomparable with other measures such as the frequently used Kendall's *tau-b*" (p. 10,582; see also Higham and Higham 2019; Masson and Rotello 2009). Other researchers (e.g., Freeman 1986; Gonzalez and Nelson 1996; Metsämuuronen 2021) propose that there is no inflation per se in *G* but, instead, a different logic of using tied pairs when computing probability. This matter is discussed later with formulae. Partly, the apparent inflation may be caused by the hidden directional nature of *G* discussed in this article.

Based on simulation results, Metsämuuronen (2021) has collected some advances of *G* in the measurement modelling settings in comparison with item–total correlation (*Rit*), item–rest correlation (*Rir*), polychoric correlation (*RPC*) and *D*. First, *G* reaches the extreme values − 1, 0, and + 1 accurately, while *Rit* and *Rir* cannot reach the extremes of correlation, and *RPC* can reach the extreme values only approximatively. Because of being based on ranks, *G* is also more robust for extreme observations, nonlinearity, and difficulty levels of the item than *Rit* and *Rir*. Hence, with binary items, *G* tends to produce estimates that underestimate item discrimination power less than the estimates by *Rit* and *Rir*. Also, *G* is applicable and accurate also with non-normal, sparse, or small datasets and crosstables, while the applicability and accuracy of the estimation result of the *Rit* and *Rir* depend on the number of categories in the variables. Second, *G* (as well as *RPC*) is accurate in reflecting the latent perfect association between the item and the score unlike *Rit*, *Rir,* and *D*; the latter behave unpredictable and they underestimate the latent perfect association in an obvious manner. While both *G* and *RPC* reflect accurately the perfect latent association, the calculation of *RPC* requires complex procedures and specific software packages while *G* is reasonably easy to calculate, even manually, in practical test settings. Also, while *RPC* refers to unknown, unreachable, and hypothetical variables that are difficult to use in further research, *G* utilizes the known composite of items and score. Many of these advances are related to SME; in comparison with PMC, both *G* and *RPC* appeared to be resistant to many sources of SME (Metsämuuronen 2021). We may add here also the result from this article: *G* has a logical directional nature from the measurement modelling viewpoint; it indicates how well the latent trait (score) explains the responses in the test items. Newson (2002) also points that the interpretation of *G* is straightforward, and it may be easier to interpret in words than PMC.

1.3 An empirical note on the identity of G and D

Traditionally, *G* is taken as a symmetric measure because it produces only one value (e.g., IBM 2017; Sheskin 2011; Sirkin 2006; Wholey et al. 2015) while *D* is unambiguously a directional measure producing three options: a symmetric estimate and two directional estimates where either of the variables is dependent and the other is independent. The latter directions are usually named as "row dependent" and "column dependent" related to the analysis of two-way contingency tables. Hence, *G* and *D* are, fundamentally, different estimators of association. However, it is easy to produce a pair of variables where the estimates by *G* and one of the directions of *D* are identical—the only requirement is that one of the variables do not have tied cases (see later Table 1).

An unpublished empirical note of the identity of *G* and *D* was made when reanalysing the published dataset by Metsämuuronen (2020a); the original analysis did not concern *G*. When reanalysing the dataset using *G*, with all variables, the estimates by *G* and a specific direction of *D* were identical. If the empirical dataset shows the identity, it can be derived also in an algebraic manner. This article shows this identity.

1.4 Research question

When knowing that, under certain conditions, $G = D \leq 1$, a relevant question is, which of the options of Somers' *D* is *G*: "row-dependent" or "column-dependent" or "symmetric"? In what follows, the forms of *G* and *D* are presented first. By comparing the formulae, it is also shown that, under certain conditions, both *G* and *D* are related to Jonckheere–Terpstra test statistic. Then, algebraic reasons for the direction of *G* are discussed. Finally, numerical examples of *G* and different varieties of *D* are given using a simulation with real-world datasets.

2 Forms of *G* and *D*

2.1 Measurement model latent to gamma and delta

The basic results in the article are general and applicable to any two general variables with ordinal or interval scale and, then, *g* and *X* refer to the variable with the narrower and wider scale, respectively. However, the applications in the article are

Table 1 Example of the estimates by G and D under different conditions; X in column

Test-taker ID	Rows (items)						Column
	$A1$	$A2$	$A3$	$B1$	$B2$	$B3$	X (score)
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	1	0	5
6	0	0	1	0	0	1	6
7	0	0	0	0	0	0	6
8	0	0	0	0	0	0	6
9	0	0	1	0	0	0	6
10	0	0	0	0	0	0	10
11	0	0	0	0	0	0	11
12	0	0	1	0	1	1	12
13	0	0	0	0	0	0	13
14	0	1	0	0	0	0	14
15	0	0	1	1	2	2	15
16	1	0	0	1	0	0	16
17	1	1	0	1	0	0	17
18	1	1	1	2	1	1	18
19	1	1	0	3	3	3	19
20	1	1	0	4	4	4	20
P	150	146	86	192	158	120	
Q	0	4	56	0	34	64	
D_R	150	150	150	192	192	192	
D_C	368	368	368	368	368	368	
$2T_R=D_R–P–Q$	0	0	8	0	0	8	
$2T_C=D_C–P–Q$	218	218	226	176	176	184	
D "column explains row"	1	0.947	0.200	1	0.646	0.292	
D "symmetric"	0.579	0.548	0.116	0.686	0.443	0.479	
D "row explains column"	0.408	0.386	0.082	0.522	0.337	0.152	
G	1	0.947	0.211	1	0.646	0.304	
tau-b	0.638	0.604	0.128	0.722	0.466	0.211	

discussed within the measurement modelling settings where the variables (item g and score or measurement scale X) are dependent because both are related to the common latent trait (θ).

Assume that the observed values in g with $r=1, \ldots, R$ and X with $c=1, \ldots, C$ distinctive ordinal or interval categories, and $R<<C$, share the common latent trait

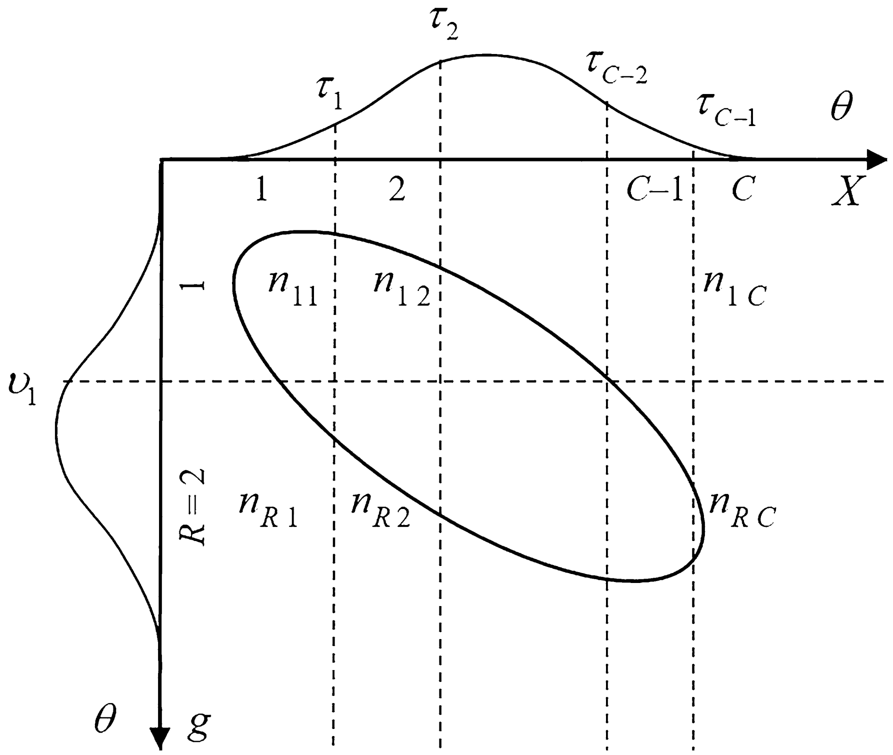

Fig. 1 A latent variable θ manifested in two ordinal variables g and X

(θ).[1] Hence, the higher the latent trait is the more probable it is to reach higher score (X) and, simultaneously, more probably a higher value (or the correct answer) in a test item (g). The threshold values of θ for each category in g are denoted by υ_i and for each category in X by τ_j. Then, g and X are related to θ so that observed value of the item is $g=x_i$, if $\upsilon_{i-1}\leq\theta<\upsilon_i$, $i=1, 2,\ldots, R$ and the observed value of the score $X=y_j$, if $\tau_{j-1}\leq\theta<\tau_j$, $j=1, 2, \ldots, C$, and $\upsilon_0 = \tau_0 = -\infty$ and $\upsilon_R = \tau_C = +\infty$. Figure 1 illustrates the model with a binary g ($R=2$); n_{ij} refers to the number of cases in in cell i, j.

2.2 Population form of γ

G estimates the probability γ that two randomly chosen cases have the same order in two variables (e.g., Van der Ark and Van Aert 2015). Let variables g and X be

[1] Notably, this is an obvious simplification of the situation. In the real-life settings, several factors and latent variables are related to the item responses such as general intelligence, reading ability, and perseverance. For the modelling purposes the one-factor model is, however, a widely used conceptualization (see e.g. Cheng et al. 2012; McDonald 1985).

sampled jointly from a bivariate population with the joint distribution π. The joint probabilities are denoted by π_{rs}. Let (x_1, y_1), (x_2, y_2), …, (x_N, y_N) be a set of observations of the joint random variables g and X. The pairs of observation (x_l, y_l) and (x_h, y_h), where $l<h$, are concordant if the order for both elements agree, that is, if $x_l<x_h$ and $y_l<y_h$ or $x_l>x_h$ and $y_l>y_h$. Similarly, the pairs are discordant when $x_l<x_h$ and $y_l>y_h$ or $x_l>x_h$ and $y_l<y_h$ simultaneously. If $x_l=x_h$ or $y_l=y_{h,}$ the pairs are tied; those are neither discordant nor concordant.

The probability that two randomly chosen test-takers have the same order in both g and X is denoted by π_P (from the traditional symbol for concordant pairs P) and the probability that two randomly chosen test-takers have a different order in g and X is denoted by π_Q (from the traditional symbol for discordant pairs Q), and

$$\pi_P = \sum_{r=1}^{g} \sum_{c=1}^{X} \pi_{rc} \left(\sum_{i>r} \sum_{j>c} \pi_{ij} + \sum_{i<r} \sum_{j<c} \pi_{ij} \right) \tag{1}$$

and

$$\pi_Q = \sum_{r=1}^{g} \sum_{c=1}^{X} \pi_{rc} \left(\sum_{i>r} \sum_{j<c} \pi_{ij} + \sum_{i<r} \sum_{j>c} \pi_{ij} \right) \tag{2}$$

(Van der Ark and Van Aert 2015). Using these symbols, the latent γ is defined as

$$\gamma = \frac{\pi_P - \pi_Q}{\pi_P + \pi_Q}. \tag{3}$$

2.3 Sample forms of *G*, *D*, and *tau-b*

The sample forms of G and D are usually expressed using the concepts of concordance (P; the number of pairs of observations into the same direction) and discordance (Q; the number of pairs into the opposite directions) observed in variables g and X. We define

$$\begin{aligned}
C_{ij} &= \sum_{h<i} \sum_{k<j} n_{hk} + \sum_{h>i} \sum_{k>j} n_{hk}, \\
D_{ij} &= \sum_{h<i} \sum_{k>j} n_{hk} + \sum_{h>i} \sum_{k<j} n_{hk}, \\
P &= \sum_{i,j} n_{ij} C_{ij}, \\
Q &= \sum_{i,j} n_{ij} D_{ij},
\end{aligned} \tag{4}$$

where n_{ij} is the number of cases in the cell ij of the two-way contingency table. Both P and Q include (the same) tied pairs; the number of these tied pairs is denoted respectively by T_g and T_X when g and X are considered. The number of all combinations of pairs related in the direction that "g given X"[2] is

$$D_r = D_g = N^2 - \sum_{i=1}^{R} \left(n_i^2\right) = \left(P + T_g\right) + \left(Q + T_g\right) = P + Q + 2T_g \tag{5}$$

and for "X given g",

$$D_c = D_X = N^2 - \sum_{i=1}^{C} \left(n_i^2\right) = \left(P + T_X\right) + \left(Q + T_X\right) = P + Q + 2T_X. \tag{6}$$

The quantities of P and Q in Eq. (4) are double of those we usually see in the textbooks (e.g., Metsämuuronen 2017; Siegel and Castellan 1988). Although calculating P and Q in practical settings is easier when only half of the directions (and then doubling them) are considered (see later Table and related discussion), the notation in Eq. (4) makes it possible to estimate the asymptotic standard errors strictly (e.g. Agresti 2010; Goodman and Kruskal 1979; Metsämuuronen 2021; see also Appendix).

The sample form of G estimates the latent γ and proportions P–Q with those pairs for which we know the direction and hence, the tied pairs are excluded:

$$G = \frac{P - Q}{P + Q}. \tag{7}$$

The asymptotic standard error (ASE) used when computing the confidence interval is

$$ASE_1(G) = \frac{4}{(P + Q)^2} \sqrt{\sum_{i,j} n_{ij}\left(QC_{ij} - PD_{ij}\right)^2} \tag{8}$$

and, under the hypotheses of independence used when computing the test statistics,

[2] See the discussion and examples of the possible confusion in naming the directions in Metsämuuronen (2020a). In the measurement modelling settings, the direction where the latent trait (score) explains the behaviour in the item is the meaningful direction (e.g., Byrne 2001). In the traditional settings of conditions, this direction is verbalized as "g given X", that is, "g is dependent on X", that is, "g dependent", and notated as $(g|X)$. However, within the traditional notation related to Somers' D, when g is "dependent," it is notated as $D(X|g)$ (e.g., Metsämuuronen 2017; Newson 2002, 2006, 2008; Siegel and Castellan 1988) inherited from the logic familiar from the general linear modelling where g is thought as independent and X as dependent. Here, the former logic familiar also from the manual calculation of Mann–Whitney U-test and Jonckheere–Terpstra test statistics is used where the dataset is first ordered by X after which the order in g is analysed, that is, the order in g depends on X. In this article, the specific notation $D(g|X)$ refers to "g dependent" or "g given X" in the spirit of conditions which, in the outputs of some generally known software packages such as IBM SPSS, SAS, as well as R libraries, would be called "X dependent." See also Table 1 and the related discussion.

$$ASE_0(G) = \frac{2}{(P+Q)}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2} \tag{9}$$

(e.g., IBM 2017; Agresti 2010) where P, Q, C_{ij}, and D_{ij} are as defined in Eq. (4).

The sample form of D proportions $P–Q$ with all possible pairs including also the tied pairs related to g or X, depending on the direction. In the case that X explains the order in g, that is, "g given X"

$$D(g|X) = \frac{P-Q}{D_g} = \frac{P-Q}{P+Q+2T_g}, \tag{10}$$

and in the case that g explains the order in X, that is, "X given g",

$$D(X|g) = \frac{P-Q}{D_X} = \frac{P-Q}{P+Q+2T_X}, \tag{11}$$

and generally, $T_g \neq T_X$. The sample form of the symmetric form of D is

$$D(sym) = \frac{P-Q}{\frac{1}{2}\left(D_g + D_X\right)} = \frac{P-Q}{P+Q+T_g+T_X}. \tag{12}$$

When computing the confidence intervals, the ASEs for $D(g|X)$ and $D(X|g)$ are

$$\mathrm{ASE}_1(D(g|X)) = \frac{2}{D_g^2}\sqrt{\sum_{i,j} n_{ij}\left(D_g\left(C_{ij} - D_{ij}\right) - (P-Q)\left(N - n_i\right)\right)^2} \tag{13}$$

and

$$\mathrm{ASE}_1(D(X|g)) = \frac{2}{D_X^2}\sqrt{\sum_{i,j} n_{ij}\left(D_X\left(C_{ij} - D_{ij}\right) - (P-Q)\left(N - n_j\right)\right)^2}. \tag{14}$$

The corresponding ASEs under the hypotheses of independence are

$$\mathrm{ASE}_0(D(g|X)) = \frac{2}{D_g}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2} \tag{15}$$

and

$$\mathrm{ASE}_0(D(X|g)) = \frac{2}{D_X}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2}. \tag{16}$$

The form of the standard error of $D(sym)$ is notably more complicated (see, e.g., IBM 2017) and it is not relevant for the latter part of the article. Hence, it is omitted here.

As a benchmark for G and D, the sample form of *tau-b* is

$$tau-b=\frac{P-Q}{\sqrt{D_g\times D_X}}=\frac{P-Q}{\sqrt{(P+Q)^2+2(P+Q)\left(T_g+T_X\right)+4\left(T_g\times T_X\right)}},\tag{17}$$

where we see that the lower magnitudes of the estimates by *tau-b* as well as $D(sym)$ in comparison with G and directional Ds are expected because *tau-b* and $D(sym)$ use the number of tied pairs in a rather extensive manner.

By comparing Eqs. (7), (10), (11), and (12) it is obvious that G gives us a more liberal approximation of the probability in comparison with D. Both ways of thinking the probability make sense. On the one hand, the logic in D is solid when we think it from the viewpoint of classic probability: the favourable cases are portioned with all cases (of pairs). On the other hand, the logic in G is the same as in the sign test (traced to Arbuthnott 1710; see Metsämuuronen 2017), and Wilcoxon signed-rank test (Wilcoxon 1945) where the sample size (related to the pairs) is adjusted by omitting the pairs where we do not know the direction. Hence, "this property [to restrict the calculation only to untied pairs in G] is neither a flaw nor a weakness" as pointed by Freeman (1986, p. 63; see also Gonzalez and Nelson 1996; Metsämuuronen 2021).

Notably, the discussion of the tied pairs can be more elaborated than above. Gonzalez and Nelson (1996), for example, separate the variables A and B as a predictor (p) and criterion (c) variables and, consequently, the tied pairs may be related to either on the predictor variable (A_p or B_p) with the number of paired cases designated as T_p, on the criterion variable (A_c or B_c) designated as T_c, or on both the predictor and the criterion variable designated as T_pc. Using these symbols, according to Gonzalez and Nelson (see also Freeman 1986), Somers' $D=(P-Q)/\left(P+Q+T_c\right)$, Kim's $D_{X.g}=(P-Q)/\left(P+Q+T_p\right)$, and Wilson's $e=(P-Q)/\left(P+Q+T_c+T_p\right)$ (see more estimators in Freeman 1986). Notably, Gonzalez and Nelson as well as Freeman simplify the set of estimators remarkably; factually, $(P-Q)/\left(P+Q+T_c\right)=D(X|g)$, Kim's $D_{X.g}=D(g|X)$, and Wilson's $e=D(sym)$. The factual directionality in G shown below is not related to the position of the variables as a predictor or criterion variable but to the widths of the scales. Hence, this logic of notation by Freeman (1986) and Gonzalez and Nelson (1996) is not used in this article.

In what follows, the sample forms and the interpretation of G and D are discussed within the measurement modelling settings and their connection to Jonckheere–Terpstra test statistic and identity under certain conditions is noted.

3 Identity of *G* and *D* and their connection to Jonckheere–Terpstra test statistic

3.1 Jonckheere–Terpstra test statistic and rank–polyserial correlation

Cureton's rank–biserial correlation coefficient (ρ_{RB}; Cureton 1956; Wendt 1972) for the association between a binary item and ordinal score can be expressed using the Mann–Whitney U test statistic (Mann and Whitney 1947):

$$\rho_{RB} = 2 \times \frac{U_{gX}^{\mathrm{Obs}}}{U_{gX}^{\mathrm{Max}}} - 1 = 2 \times \frac{U_1}{n_0 n_1} - 1, \tag{18}$$

where U_{gX}^{Obs} is the observed U test value related to the higher[3] of the subsamples ($l=0$ and $h=1$) in g, U_{gX}^{Max} is the theoretical maximum value of U test, and n_0 and n_1 are the numbers of cases in the subsamples in g. $U_{gX}^{Max} = n_0 n_1$ implies the condition that all test-takers in the higher subsample $h=1$ are ranked higher in X than the test-takers in the lower subsample $l=0$.

Jonckheere–Terpstra test statistic (*JT;* Jonckheere 1954; Terpstra 1952) extends the directional U and its calculation procedure to polytomous cases (e.g., Metsämuuronen 2017; Siegel and Castellan 1988). Hence, logically, the following measure may be called rank–*poly*serial correlation (ρ_{RP}):

$$\rho_{RP} = 2 \times \frac{JT_{gX}^{\mathrm{Obs}}}{JT_{gX}^{\mathrm{Max}}} - 1 = 2 \times \frac{JT}{\sum\limits_{l<h}^{R} n_l n_h} - 1, \tag{19}$$

where JT_{gX}^{Obs} and JT_{gX}^{Max} are the observed and maximal JT statistic. The characteristics of the measure are not discussed here although we note that JT statistic is embedded in the core of the measure. The core in ρ_{RP} is the probability measure $JT \Big/ \sum\limits_{l<h}^{R} n_l n_h$ ranging 0–1 and indicating the proportion of logically ordered observations in g after they are ordered by X. In Eq. (19), this measure is transformed, using a linear transformation of doubling and centring, to the same scale as the correlation ranging –1 to +1. With a binary g, ρ_{RB} is a special case of ρ_{RP}.

3.2 Relation of *D* and JT

Consider the direction of conditions where "g given X". Because of Eq. (5)

$$P + Q = \left(N^2 - \sum_{i=1}^{R} \left(n_i^2 \right) \right) - 2T_g, \tag{20}$$

The number of cases in the subsamples related to g and X are n_i and n_j, respectively, and, then

$$N = \sum_{i=1}^{R} n_i = \sum_{j=1}^{C} n_j. \tag{21}$$

Because of (21), the element $N^2 - \sum\limits_{i=1}^{R} \left(n_i^2 \right)$ can be manipulated as follows:

[3] Cf. Wendt's (1972) modification where U is based on the *lower* of subsamples $l=0$.

$$N^2 - \sum_{i=1}^{R}\left(n_i^2\right) = \left(\sum_{i=1}^{R} n_i\right)^2 - \sum_{i=1}^{R}\left(n_i^2\right) = \sum_{i=1}^{R}\left(n_i^2\right) + 2 \times \sum_{l<h}^{R} n_l n_h - \sum_{i=1}^{R}\left(n_i^2\right) = 2\sum_{l<h}^{R} n_l n_h. \tag{22}$$

Hence, because of (10) and (22), $D(g|X)$ can be rewritten as

$$D(g|X) = \frac{P-Q}{N^2 - \sum_{i=1}^{R}\left(n_i^2\right)} = \frac{P-Q}{2\sum_{l<h}^{R} n_l n_h}. \tag{23}$$

Because of Eqs. (20) and (22)

$$Q = 2\sum_{l<h}^{R} n_l n_h - \left(P + 2T_g\right). \tag{24}$$

Then, $D(g|X)$ can be rewritten as

$$\begin{aligned} D(g|X) &= \frac{P-Q}{2\sum_{l<h}^{R} n_l n_h} = \frac{P - \left(2\sum_{l<h}^{R} n_l n_h - \left(P + 2T_g\right)\right)}{2\sum_{l<h}^{R} n_l n_h} \\ &= \frac{2\left(P + T_g\right) - 2\sum_{l<h}^{R} n_l n_h}{2\sum_{l<h}^{R} n_l n_h} = \frac{\left(P + T_g\right)}{\sum_{l<h}^{R} n_l n_h} - 1. \end{aligned} \tag{25}$$

Because of the definition in Eq. (5), including both positive and negative direction, the element $\left(P + T_g\right)$ is two times the number of pairs in one direction. Remembering that JT equals the number of pairs in the same order including only the positive elements,

$$P + T_g = 2 \times JT. \tag{26}$$

Then, because of Eqs. (23), (26), and (19), we note the identity of ρ_{RP} and D:

$$D(g|X) = \frac{P-Q}{\sum_{l<h}^{R} n_l n_h} = 2 \times \frac{JT}{\sum_{l<h}^{R} n_l n_h} - 1 = \rho_{RP}, \tag{27}$$

that is,ρ_{RP} is a special case of Somers' D so directed that "g given X". Hence, in measurement modelling settings, Somers' $D(g|X)$ strictly indicates the proportion of logically ordered tests-takers in the item after they are ordered by the score.

3.3 Relation of *G* and JT

Because of Eqs. (7) and (20)

$$G = \frac{P-Q}{P+Q} = \frac{P-Q}{\left(N^2 - \sum_{i=1}^{R}\left(n_i^2\right)\right) - 2T_g} = \frac{P-Q}{2\sum_{l<h}^{R} n_l n_h - 2T_g}. \tag{28}$$

Because of Eqs. (28) and (24), parallel to Eq. (25), we can write

$$G = \frac{P-Q}{2\sum_{l<h}^{R} n_l n_h - 2T_g} = \frac{P - 2\sum_{l<h}^{R} n_l n_h + P + 2T_g}{2\sum_{l<h}^{R} n_l n_h - 2T_g} = \frac{P+T_g}{\sum_{l<h}^{R} n_l n_h - T_g} - 1 \tag{29}$$

and because of Eq. (26)

$$G = \frac{P-Q}{P+Q} = 2 \times \frac{JT}{\sum_{l<h}^{R} n_l n_h - T_g} - 1. \tag{30}$$

This indicates that, in the measurement modelling settings, Goodman–Kruskal gamma can be interpreted as a slightly modified proportion of the logically ordered tests-takers in the item after they are ordered by the score while taking into account only those cases for which we know the order, that is, considering only the pairs without ties.

While the coefficient in Eq. (19) is called the rank–polyserial correlation coefficient, also the latter part in Eq. (30) could be used as ρ_{RP}. However, the former estimator related to *D* (Eq. 27) gives a more conservative estimate while the latter related to *G* (Eq. 30) gives a more liberal estimate of the association between two ordinal-scaled variables.

3.4 Identity of *G* and *D*

Strictly from Eqs. (7), (10), (11), and (12) it is known that $G = D$ when there are no tied pairs. Then, *G* has the identity of *D* under three general conditions irrespective of the distributions in the variables, difficulty level in variables, number of cases, number of categories in the variables, and number of ties in the single variables: (1) when either of the variables is or both are continuous, implying no tied pairs; (2) if *X* is not continuous but there are no ties in *X*, that is, when each test-taker gets unique score regardless the distribution in the item, and (3) when there are ties in *X* but there are no crossing observations between *g* and *X*, that is, when all the tied values in the score are related to the identical value in item. The last of the options appears to be important in understanding the direction in *G*.

From the direction of "g given X", when $T_g = 0$, because $T_g \neq T_{\mathrm{X},}$ and because of Eqs. (7) and (10),

$$D(g|X) = \frac{P-Q}{P+Q+2T_g} = \frac{P-Q}{P+Q} = G. \tag{31}$$

Similarly, from the direction of "X given g", when $T_X = 0$, because of Eqs. (7) and (11)

$$D(X|g) = \frac{P-Q}{P+Q+2T_X} = \frac{P-Q}{P+Q} = G. \tag{32}$$

However, the condition of $T_X = 0$ is possible only in the case of continuous variables causing $T_g = T_X = 0$ and, then $G = D(X|g) = D(g|X) = D(sym)$. The reason is that, excluding the continuous case, the condition of $T_g = 0$ or $T_X = 0$ is true only when there are ties in the variables but there are no crossing observations between the two variables, that is, when all the tied values in one variable are related to an identical value in the other variable (see variables A2 and B2 in Table 1). This can happen only with the variable that has a wider scale because, in the variable with a shorter scale, there will always be at least two pairs that are tied with the variable with a wider scale. Hence, only the variable with the wider scale can be the one causing the condition of no ties ($T_g = 0$) irrespective of whether the variable is in row or in column or whether it is a predictor or criterion variable (see Gonzalez and Nelson 1996). Therefore, except the case of continuous variables implying $T_g = T_X = 0$, when $T_g = 0$ and $T_X \neq 0$,

$$\begin{aligned} G &= \frac{P-Q}{P+Q} = \frac{P-Q}{P+Q+T_g} \\ &= D(g|X) = G(g|X) \neq D(X|g). \end{aligned} \tag{33}$$

Equation (33) means that, although usually taken as a symmetric measure, Goodman–Kruskal gamma is, in fact, a directional measure the same manner as is Somers D; G is directed so that the order in the variable with the wider scale explains the order in the variable with the narrower scale without the relation to rows and columns in the cross-tables. Numerical examples will clarify the phenomenon.

Notably, also, except the case of continuous variables when $T_X = T_g = 0$, the ASEs of G and D are equal only in one direction. This is easy to show for the ASEs under the hypotheses of independency. Because of Eq. (5) and (6), $D_g = P + Q + 2T_g \neq P + Q + 2T_X = D_X$. Because of Eqs. (9) and (15), knowing that $T_X = 0$ can be obtained only with continuous variables without tied pairs, under any other condition when $T_g = 0 \neq T_X$,

$$\begin{aligned} ASE_0(G) &= \frac{2}{(P+Q)}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij}-D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2}, \\ &= \frac{2}{(P+Q+0)}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij}-D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2}, \\ &= ASE_0(D(g|X)) \neq ASE_0(D(X|g)) \end{aligned} \tag{34}$$

Also, the empirical findings (see discussion with Table 2 below) suggest that under the same conditions as above, $ASE_1(G) = ASE_1(D(g|X)) \neq ASE_1(D(X|g))$ although showing this algebraically is not obvious; the formulae (8), (13), and (14) use different sources of information.

4 Numerical examples

4.1 A simple comparison

The estimates by G and D are compared first using a simple dataset with two sets of variables with a narrower scale (Table 1): a binary set (items $A1$, $A2$, $A3$) and a polytomous set (items $B1$, $B2$, $B3$). In both sets, one item follows a deterministic pattern without tied pairs and without stochastic error ($A1$ and $B1$)—here we expect to see perfect item discrimination and $D = G = 1$; one item without tied pairs and including stochastic error ($A2$ and $B2$)—here we expect to see $G = D \leq 1$; and one item with tied pairs and including stochastic error ($A3$ and $B3$)—here we expect to see $G > D$. In all cases, the score with the wider scale includes a small number of tied cases, just to show their effect in the estimates. As an example, the estimates and statistics by variable $g = A2$ and X are illustrated in its form of two-ways contingency table (Table 2).

Given Table 2, the number of pairs in the same direction is $P = 2 \times (13 \times 5 + 2 \times 4) = 146$ and the number of pairs in the opposite directions is $Q = 2 \times (13 \times 0 + 2 \times 1) = 4$, and consequently, $P - Q = 142$ and $P + Q = 150$. The number of all pairs in the direction of "g given X" is $D_g = 20^2 - (225 + 25) = 150$ and the number of pairs in the direction of "X given g" is $D_X = 20^2 - (1 + 1 + ... + 1) = 400 - 32 = 368$. Then $G = 142/150 = 0.947$,

Table 2 Two-way contingency table for variables $A2$ and X related to Table 1

		X																		
		1	2	3	4	5	6	10	11	12	13	14	15	16	17	18	19	20	Total	n_i^2
$A2$	0	1	1	1	1	1	4	1	1	1	1	0	1	1	0	0	0	0	15	225
	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	5	25
Total		1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	20	
n_j^2		1	1	1	1	1	16	1	1	1	1	1	1	1	1	1	1	1		

$D(g|X) = 142/150 = 0.947$, $D(X|g) = 142/368 = 0.386$, $D(sym) = 142\Big/\frac{1}{2}(150 + 368) = 0.548$, and $tau - b = 142\Big/\sqrt{(150 \times 368)} = 0.604$.

The calculation of the ASEs and confidence intervals of G and D are presented in Appendix. Given Table 2, $\mathrm{ASE}_1(G) = 0.0592$ and $\mathrm{ASE}_0(G) = 0.2515$. Then, the traditional asymptotic 95% confidence interval for the true γ is $\gamma = 0.947 \pm 2.101 \times 0.0592 = [0.822, 1]$[4] and the asymptotic significance, when testing the hypothesis $\gamma{=}0$, is $Z = 0.947/0.2515 = 3.764$ leading to $p{<}0.001$. The corresponding ASEs of the directed Ds are $ASE_1(D(g|X)) = 0.0592$, $\mathrm{ASE}_0(D(g|X)) = 0.2515$, $ASE_1(D(X|g)) = 0.1003$, and $\mathrm{ASE}_0(D(g|X)) = 0.1025$ (see Appendix). Somers' D estimates the true probability δ. Then, the traditional asymptotic 95% confidence intervals for δ are $\delta(g|X) = 0.947 \pm 2.101 \times 0.0592 = [0.822, 1]$, and $\delta(X|g) = 0.386 \pm 2.101 \times 0.1003$= [0.261, 0.597]. When testing the hypothesis $\delta(g|X){=}0$, $Z = 0.947/0.2515 = 3.764$ with $p{<}0.001$ and for $\delta(X|g){=}0$, $Z = 0.3859/0.1025 = 3.764$ with $p{<}0.001$. We note the identical test statistics and identical statistical inference by D as with G.

Some lifts from Tables 1 and 2 are highlighted. First, we note the relevant direction of association discussed in Footnote 2. G ("g given X") and D ("g given X") are the estimators that detect the deterministic pattern of item discrimination in items A1 and B1. This was expected because of Eqs. (19), (27) and (30) related to JT statistic; in the deterministic patterns as in A1 and B1, $JT_{gX}^{Obs}{=}JT_{gX}^{Max}$ and, consequently, $G{=}D{=}1$. Second, with items A1 and A2 as well as B1 and B2, $G = D(g|X) \neq D(X|g) \neq D(Sym)$ because there are no tied pairs related to X and, then, $(P{+}Q){=}N^2 - \sum_{i=1}^{R}\left(n_i^2\right) \neq N^2 - \sum_{j=1}^{C}\left(n_j^2\right)$. This was expected because of Eq. (33). Third, when there are tied pairs (A3 and B3), $G > D(g|X)$ because $P{+}Q{<}P + Q + 2T_g$. This is expected because of Eqs. (7) and (10). Fourth, in the case that there are no tied pairs (A1, A2, B1, B2), $ASE_1(G) = ASE_1(D(g|X)) \neq ASE_1(D(X|g))$ and $ASE_0(G) = ASE_0(D(g|X)) \neq ASE_0(D(X|g))$. This is expected because of Eq. (34).

To verify the result, we could restudy the dataset by pivoting the cross-tables such that X is the row factor and the g is column factor. We would see that, in items $A1, A2, B1$, and $B2$, $G(g|X){=}D(g|X){\neq}D(X|g)$.

4.2 A comparison of *G* and *D* with a larger dataset

In the second comparison of the estimates by G and D, a wider survey of the behaviour of G and the difference variants of D was conducted. In the comparison, 13,392 test items from 1,292 tests were formed by different combinations of single items and sub-scores constructed by different item compilations based on randomly

[4] Notably, the upper limit is truncated to 1 because γ, as being probability, cannot exceed 1. This logic corresponds with the logic traditionally used with corrected coefficients for eta squared (i.e., epsilon squared and omega squared); when the squared values get out-of-range values below zero, these are traditionally truncated to be 0 (see, e.g., Cohen 1973; Okada 2017). Another option with G, pointed by an anonymous reader, would be a variable transformation.

Table 3 Average estimates of *G* and *D* based on the real-world datasets

df(*g*)	Mean				Std. deviation				*N*
	G	*D*(*g*|*X*)	*D*(*sym*)	*D*(*X*|*g*)	*G*	*D*(*g*|*X*)	*D*(*sym*)	*D*(*X*|*g*)	
1	0.650	0.628	0.388	0.284	0.144	0.144	0.104	0.084	7131
2	0.694	0.670	0.522	0.428	0.103	0.104	0.090	0.082	2715
3	0.729	0.704	0.596	0.518	0.085	0.086	0.080	0.077	1233
4	0.765	0.737	0.655	0.590	0.070	0.072	0.071	0.071	656
5	0.782	0.753	0.690	0.636	0.063	0.064	0.064	0.063	413
6	0.806	0.778	0.724	0.677	0.050	0.051	0.054	0.056	336
7	0.830	0.799	0.756	0.717	0.046	0.048	0.050	0.052	231
8	0.844	0.812	0.775	0.741	0.041	0.044	0.046	0.048	118
9	0.865	0.835	0.805	0.778	0.043	0.045	0.047	0.049	161
10	0.880	0.847	0.823	0.800	0.034	0.036	0.037	0.039	140
11	0.894	0.861	0.840	0.820	0.025	0.030	0.031	0.033	98
12	0.898	0.869	0.850	0.832	0.026	0.032	0.034	0.036	82
13–15	0.893	0.864	0.847	0.830	0.021	0.026	0.028	0.029	78
Total	0.694	0.670	0.493	0.404	0.135	0.134	0.161	0.173	13,392

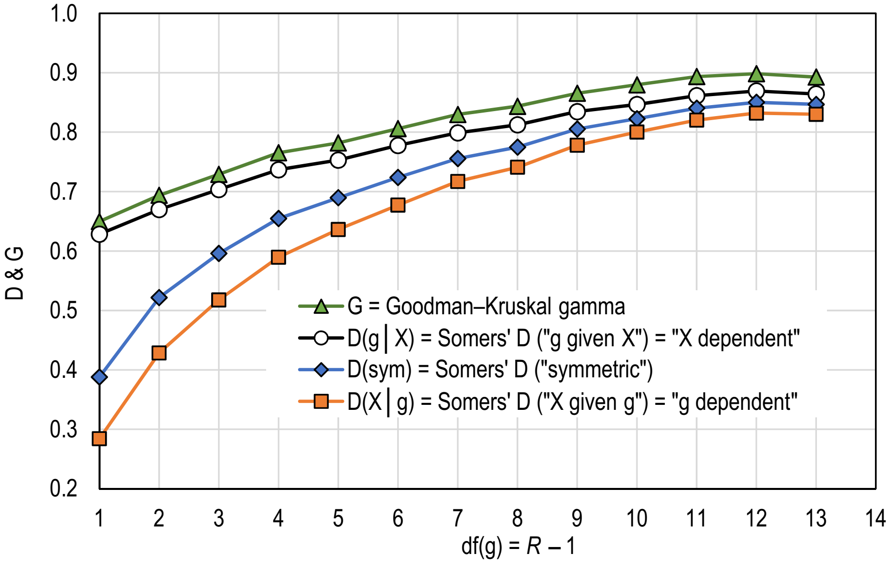

Fig. 2 Comparison of G and D by the degrees of freedom of the item; *df*(*g*) = *R*–1; *df*(*g*) = 13 is combined 13–15; *k* = 13,392 items

selected test-takers from a national-level dataset of 4,000 test-takers of a mathematics test for grade 9 with 30 binary items (FINEEC 2018). In the original dataset, the item discrimination ranged 0.332 < $PMC = \rho_{gX} = Rit$ < 0.627 with the average

$\overline{Rit} = 0.481$, the difficulty levels of the items ranged $0.24 < p < 0.95$ with the average difficulty level of $\overline{p} = 0.63$, and with the lower bound of reliability of $\alpha = 0.885$. A small number of artificial datasets (13% of tests) were constructed to cover the very difficult and extremely difficult tests. Finally, a set of 1,292, mostly real-world datasets with different number of test-takers ($N = 50$–100–200), test lengths ($k = 2$–30), difficulty levels ($\overline{p} = 0.08$–0.96), reliabilities ($\alpha = 0.74$–0.98), and degrees of freedom in the item $df(g) = 1$–15, and in the score $df(X) = 12$–27) with 13,392 partly related test items was formed to compare the estimates by G and D. The average estimates are collected in Table 3 and Fig. 2. The main outcome of the survey is that G really follows the trend of $D(g \mid X)$ and not $D(X \mid g)$ nor the symmetric D (Fig. 2). Using the same logic of naming as with D, G is, factually, $G(g \mid X)$.

5 Conclusions and possibilities of *G* in the measurement modelling settings

5.1 General notes on the results

The main result is that, although Goodman–Kruskal gamma is usually taken as a symmetric measure, it is, in fact, a *directional measure* the same manner as is Somers' D. The direction in G is not determined by rows and columns but, instead, G is directed to the way where the order in the variable with a narrower scale depends on the order in the variable with a wider scale in the analysis. This direction makes sense in the measurement modelling settings where it is assumed that the latent trait manifested as the score or the measurement scale with wider scale explains the response pattern in the item with the narrower scale (see, e.g., Kim 1971; Byrne 2001; Metsämuuronen 2017). This directional nature of G may explain partly the potential "inflation" discussed by, for example, Higham and Higham (2019), Kvålseth (2017), and Masson and Rotello (2009).

That G is a directional measure is somewhat alarming from the viewpoint of using it in general settings; it is *strongly recommendable that gamma should not be used as a symmetric measure,* and it *should be used directionally only when willing to explain the response pattern in a variable with a narrower scale by the variable with a wider scale*.

5.2 Possibilities of *G* in the measurement modelling settings

That G is not related to rows and columns but to the widths of the scales is a positive matter within measurement modelling settings: G leads strictly to the logical direction from the theory viewpoint where the latent trait manifested as the score or the measurement scale drives the responses in the test item. Hence, G could be an asset in measurement modelling settings. While $D(g|X)$ is raised as one of the "superior alternatives" to PMC in the binary case (Metsämuuronen 2020a), G would be even "more superior alternative" than D (Metsämuuronen 2021). After all, while D tends to underestimate association of an item and the score in an obvious manner when the item has three categories or more (see Göktaş and İşçi 2011; Metsämuuronen 2020a), G does not underestimate IDP to that extent (Metsämuuronen 2021).

As being a directional measure and one of the "superior alternatives" to PMC, G have strict relevance in a new concept of "SME-corrected" estimates of reliability proposed by Metsämuuronen (2021). Namely, it is known that coefficient alpha, as an example, a classical estimator of the lower bound of reliability, can be expressed using item–total correlation (PMC $= \rho_{gX}$):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{g=1}^{k} \sigma_g^2}{\left(\sum_{g=1}^{k} \sigma_g \rho_{gX}\right)^2}\right) \tag{35}$$

(Lord and Novick 1968). It is also known that the element ρ_{gX} = PMC in Eq. (35) always underestimates the true association between the item and the score because of RR and several other sources of SME (Metsämuuronen 2021) and, hence, the magnitude of the estimate of reliability is reduced because of mechanical reasons. If we use G instead of PMC in the form, we will get one option for a "SME-corrected" estimate of reliability:

$$\alpha_G = \frac{k}{k-1}\left(1 - \frac{\sum_{g=1}^{k} \sigma_g^2}{\left(\sum_{g=1}^{k} \sigma_g G_{gX}\right)^2}\right) \tag{36}$$

(Metsämuuronen 2021), which suffer remarkable less loss of information related to SME than the original estimator because G is less affected on SME (Metsämuuronen 2020c, 2021). The matter is not elaborated further here; more "SME-corrected" formulae of reliability can be found in Metsämuuronen (2021) and, specifically, in Metsämuuronen (2020c). More studies in this area would enrich our knowledge of the matter.

Another possibility in the directionality of G is its potential use as an indicator of explaining power in the form of G^2 in the same manner as we use ρ_{XY}^2 for two metric variables and η^2 for a categorical and a metric variable. The advance of G^2 is that, in the case that the scales of the variables differ from each other, unlike ρ_{XY}^2 and η^2, G^2 can reach correctly also the extreme value + 1. Assumingly, using G^2 would give us a kind of "SME-corrected" estimate of the explaining power indicating how well the variable with a wider scale explains the response pattern in the variable with a narrower scale. This would be useful, specifically, in the measurement modelling settings with binary items where ρ_{XY}^2 and η^2 may underestimate the association remarkably. This area would be worth studying more.

General advances of G in the measurement modelling settings were already discussed Introduction (see also Metsämuuronen 2021).

5.3 Limitations

An obvious limitation of the study is that the survey with real-world items that was used to illustrate the connection between G and the variations of D carries its own limitations. Although the numbers of subtests ($n = 1296$) and items ($k = 13{,}392$) used in the survey are rather convincing, those are based on *one* basic dataset. Results may have been somewhat different if truly polytomous test items were used in the simulation. Replications of the design or another approach with a more independent estimates may increase our knowledge of the relation between the estimators. Another obvious limitation is that the algebraic connection of the obviously different forms of $\mathrm{ASE}_1(G)$ and $\mathrm{ASE}_1(D(g|X))$ was noted in the empirical dataset but it was not shown in an algebraic manner.

When it comes to G itself, because of carrying, largely, the same characteristics as D, G also has some of the known disadvantages noticed in D. In item analysis settings, one of these is that G tends to underestimate the association of g and X in an obvious manner, when the number of categories in g is large, more than four (Metsämuuronen 2021). Because G gives obvious underestimates of association in comparison with PMC, some correction may be proper to propose to enhance G against this deficiency. The possible correction needed in G in item analysis settings, where the score and the items are manifestations of the *same* latent variable and when we have a mechanical correction between these variables, is, undoubtedly, different than in the case of two independent variables. One option, suggested by Metsämuuronen (2021), a “dimension-corrected gamma” (G_2), specific to the measurement modelling settings, transforms the linear nature in G toward the trigonometric nature. G_2 seems to overcome the problem of obvious underestimation in item analysis settings without producing obvious overestimates (Metsämuuronen 2021). Studying these kinds of coefficients may enrich the discussion related to “SME-corrected” reliability (see above).

5.4 Further suggestions

Because G appears to be a directional measure, the developers of the enhanced or corrected G (e.g., Bai and Wei 2009; Highan and Higham 2019; Hryniewicz 2006; Kvålseth 2017; Masson and Rotello 2009; Rousson 2007) or enhanced procedures to estimate the confidence intervals for G (e.g., Van der Ark and Van Aert 2015; Woods 2007) may be willing to consider, if needed, their correction factors or estimators from this viewpoint also. Maybe the researcher working with D in connection with Harrell’s C and the related AUC and ROC (see Harrell 2001; Harrell et al. 1982; Heagerty and Zheng 2005), would be interesting in considering to study further the possibilities of G in relation with those tools (see Heagerty and Zheng 2005; Higham and Higham 2019) from the directionality viewpoint. Obviously, it would be suggested to reconsider the texts also in the textbooks and manuals considering G as a symmetric measure (e.g., IBM 2017; Metsämuuronen 2017; Sheskin 2011; Sirkin 2006; Wholey et al. 2015).

All in all, the directional nature in G and D may be worth considering within the measurement modelling settings. A relevant question arising from the directionality embedded to G and D is why are we would use, in the first place, the *non*directional correlation coefficients while the philosophy of measurement modelling is based on the idea of directionality that the latent trait manifested as the score drives to the observed behaviour in the item and not the other way round. Then, studying the family of the directional coefficients of correlation could enrich the discussions related to such areas as the estimators of reliability or item discrimination power, or calculating the factor loadings used in estimating maximal reliability. G and D with a directional nature could be worth considering in these areas.

Appendix: Calculation of the ASEs based on Table 2

Given Table 2, the sub-components for the ASEs of G are:
$\sum_{i,j} n_{ij}\left(QC_{ij} - PD_{ij}\right)^2 = 111{,}000$, $\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 = 1{,}364$, $\frac{1}{N}(P-Q)^2 = 1008.2$, $\frac{4}{(P+Q)^2} = 1.778 \times 10^{-4}$ and $\frac{2}{P+Q} = 0.0133333$.

Because of Eqs. (8) and (9),

$$\mathrm{ASE}_1(G) = 1.778 \times 10^{-4}\sqrt{111{,}000} = 0.0592$$

and

$$\mathrm{ASE}_0(G) = 0.0133333\sqrt{1{,}364 - 1{,}008.2} = 0.2515.$$

The traditional asymptotic 95% confidence interval for the true γ is $\gamma = G \pm t_{\alpha 0.975}(19) \times \mathrm{ASE}_1(G) = 0.947 \pm 2.101 \times 0.0592 = [0.822, 1]$ and the asymptotic significance, when testing the hypothesis $\gamma = 0$, is $Z = \frac{G}{\mathrm{ASE}_0(G)} = \frac{0.947}{0.2515} = 3.764$ leading to $p < 0.001$.

For the ASEs of D,

$$\sum_{i,j} n_{ij}\left(D_g\left(C_{ij} - D_{ij}\right) - (P-Q)\left(N - n_i\right)\right)^2 = 444{,}000,$$

$$\sum_{i,j} n_{ij}\left(D_X\left(C_{ij} - D_{ij}\right) - (P-Q)\left(N - n_j\right)\right)^2 = 46{,}130{,}880,$$

$$\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2 = 1{,}364 - 1{,}008.2 = 355.8,$$

$\frac{2}{D_g^2} = 8.8888 \times 10^{-5}$, and $\frac{2}{D_X^2} = 1.4768 \times 10^{-5}$,
Then, because of Eqs. (13), (14), (15), and (6),

$$\mathrm{ASE}_1(D(g|X)) = 8.8888 \times 10^{-5} \times \sqrt{444,000} = 0.0592,$$

$$\mathrm{ASE}_1(D(X|g)) = 1.4768 \times 10^{-5} \times \sqrt{46,130,880} = 0.1003,$$

$$\mathrm{ASE}_0(D(g|X)) = 0.013333 \times \sqrt{355.8} = 0.2515, \text{ and}$$

$$\mathrm{ASE}_0(D(X|g)) = 0.00543 \times \sqrt{355.8} = 0.1025.$$

The traditional asymptotic 95% confidence intervals for δ are.
$\delta(g|X) = D(g|X) \pm t_{\alpha 0.975}(19) \times ASE_1(D(g|X)) = 0.947 \pm 2.101 \times 0.0592 = [0.822,\ 1]$, and $\delta(X|g) = 0.386 \pm 2.101 \times 0.1003 = [0.261,\ 0.597]$. When testing the hypothesis $\delta(g|X){=}0$, $Z = \frac{0.947}{0.2515} = 3.764$ with $p<0.001$ and for $\delta(X|g){=}0$, $Z = \frac{0.3859}{0.1025} = 3.764$ with $p<0.001$.

Acknowledgements The writer sincerely thank Dr. Roger Newson, research associate at the Faculty of Medicine, School of Public Health, Imperial College, London, for leading the writer to find, in an early phase of studying the matter, an essential mistake in the derivations regarding the connection of *G* and *D*. In the earlier phases, he also helped in getting access to other useful resources on the topic related to Somers' *D* including the Greiner's relation that explains the reason for the underestimation of association in both *D* and *G*. Sincere thanks also to Jukka Marjanen, Counsellor of Evaluation in FINEEC, Finland, for helping in solving a practical problem in the manual calculations of the ASEs.

Declarations

Conflict of interest There is no conflict of interest.

References

Agresti A (2010) Analysis of ordinal categorical data, 2nd edn. Wiley, New Jersey

Arbuthnott J (1997) An argument for divine providence, taken from the constant regularity observed in the births of both sexes. Philos Trans R Soc Lond 27(325–336):186–190. https://doi.org/10.1098/rstl.1710.0011

Bai J, Wei L-L (2009) A new method of attribute reduction based on gamma coefficient. Proc WRI Glob Congr Intell Syst. https://doi.org/10.1109/GCIS.2009.212

Byrne BM (2001) Structural equation modelling with AMOS. Basic concepts, applications, and programming. Lawrence Erlbaum Associates Publishers, Mahwah

Cheng Y, Yuan K-H, Liu C (2012) Comparison of reliability measures under factor analysis and item response theory. Educ Psychol Meas 72(1):52–67. https://doi.org/10.1177/0013164411407315

Cohen J (1973) Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educ Psychol Meas 33(1):107–112. https://doi.org/10.1177/001316447303300111

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

Davis JA (1967) A partial coefficient for Goodman and Kruskal's gamma. J Am Stat Assoc 62(317):189–193. https://doi.org/10.1080/01621459.1967.10482900

Forthmann B, Förster N, Schütze B, Hebbecker K, Flessner J, Peters MT, Souvignier E (2020) How much *g* is in the distractor? Re-thinking item-analysis of multiple-choice items. J Intelligence 8(1):11. https://doi.org/10.3390/jintelligence8010011

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

Freeman LC (1986) Order-based statistics and monotonicity: a family of ordinal measures of association. J Math Sociol 12(1):49–69. https://doi.org/10.1080/0022250X.1986.9990004

Göktaş A, İşçi OA (2011) Comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. Metodološki Zvezki 8(1):17–37

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

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

Goodman LA, Kruskal WH (1979) Measures of association for cross classification. Springer-Verlag, Berlin

Greiner R (1909) Über das Fehlersystem der Kollektivmaßlehre (Of the error systemic of collectives). J Math Phys 57:121–158

Harrell F (2001) Regression modelling strategies. Springer

Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. Journal of the American Medical Association 247(18):2543–2546. https://doi.org/10.1001/jama.1982.03320430047030

Heagerty PJ, Zheng Y (2005) Survival model predictive accuracy and ROC curves. Biometrics 61(1):92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

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

Hryniewicz O (2006) Goodman-Kruskal γ measure of dependence for fuzzy ordered categorical data. Comput Stat Data Anal 51(1):323–334. https://doi.org/10.1016/j.csda.2006.04.014

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

Jonckheere AR (1954) A distribution-free k–sample test against ordered alternatives. Biometrika 41(1–2):133–145. https://doi.org/10.1093/biomet/41.1-2.133

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

Kendall MG (1948) Rank correlation methods, 1st edn. Charles Griffin & Co Ltd., Glasgow

Kendall MG, Gibbons JD (1990) Rank correlation methods, 5th edn. Oxford University Press, Oxford

Kim J-O (1971) Predictive measures of ordinal association. Am J Sociol 76(5):891–907

Kreiner S, Christensen KB (2009) Item screening in graphical loglinear Rasch models. Psychometrika 76(2):228–256. https://doi.org/10.1007/s11336-011-9203-y

Kvålseth TO (2017) An alternative measure of ordinal association as a value-validity correction of the Goodman-Kruskal gamma. Commun Stat Theory Methods 46(21):10582–10593. https://doi.org/10.1080/03610926.2016.1239114

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

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50–60. https://doi.org/10.1214/aoms/1177730491

Martin WS (1973) The effects of scaling on the correlation coefficient: a test of validity. J Mark Res 10(3):316–318. https://doi.org/10.2307/3149702

Martin WS (1978) Effects of scaling on the correlation coefficient: additional considerations. J Mark Res 15(2):304–308. https://doi.org/10.1177/002224377801500219

Masson MEJ, Rotello CM (2009) Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: implications for studies of metacognitive processes. J Exp Psychol Learn Mem Cogn 35(2):509–527. https://doi.org/10.1037/a0014876

McDonald RP (1985) Factor analysis and related methods. Lawrence Erlbaum Associates, New Jersey

Mendoza JL, Mumford M (1987) Corrections for attenuation and range restriction on the predictor. J Educ Stat 12(3):282–293. https://doi.org/10.3102/10769986012003282

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

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

Metsämuuronen J (2020b) Dimension-corrected Somers' D for the item analysis settings. Int J Educ Methodol 6(2):297–317. https://doi.org/10.12973/ijem.6.2.297

Metsämuuronen J (2020c) Seeking the real reliability. Rethinking the measurement model from the viewpoint of systematic mechanical error related to the estimators of association. ResearchGate. https://doi.org/10.13140/RG.2.2.10599.88484

Metsämuuronen J (2021) Goodman-Kruskal gamma and dimension-corrected gamma in educational measurement settings. Int J Educ Methodol 7(1):95–118. https://doi.org/10.12973/ijem.7.1.95

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

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

Newson R (2008) Identity of Somers' *D* and the rank biserial correlation coefficient. http://www.rogernewsonresources.org.uk/miscdocs/ranksum1.pdf. Accessed 1 Apr 2021

Nielsen T, Santiago PHR (2020) Using graphical loglinear Rasch models to investigate the construct validity of Perceived Stress Scale. In: Khine MS (ed) Rasch measurement: applications in quantitative educational research. Springer Nature, Berlin, pp 261–281. https://doi.org/10.1007/978-981-15-1800-3_14

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

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

Pearson K (1896) Mathematical contributions to the theory of evolution III. Regression, heredity, and panmixia. Philos Trans R Soc A 187:253–318. https://doi.org/10.1098/rsta.1896.0007

Pearson K (1900) I. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Philos Trans R Soc A 195(262–273):1–47. https://doi.org/10.1098/rsta.1900.0022

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

Rousson V (2007) The gamma coefficient revisited. Statist Probab Lett 77(17):1696–1704. https://doi.org/10.1016/j.spl.2007.04.009

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

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

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

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

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

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

Terpstra TJ (1952) The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. Indag Math 14(3):327–333. https://doi.org/10.1016/S1385-7258(52)50043-X

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

Wendt HW (1972) Dealing with a common problem in social science: a simplified rank biserial coefficient of correlation based on the U statistic. Eur J Soc Psychol 2(4):463–465. https://doi.org/10.1002/ejsp.2420020412

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

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometr Bull 1(6):80–83. https://doi.org/10.2307/3001968

Wilson TP (1974) Measures of association for bivariate ordinal hypotheses. In: Blalock HM (ed) Measurement in the social sciences. Macmillan Education, Aldine, pp 327–342

Woods CM (2007) Confidence intervals for gamma-family measures of ordinal association. Psychol Methods 12(2):185–204. https://doi.org/10.1037/1082-989X.12.2.185