Research Article doi: 10.12973/ijem.7.1.95

International Journal of Educational Methodology
Volume 7, Issue 1, 95 - 118.
ISSN: 2469-9632
https://www.ijem.com/

Goodman–Kruskal Gamma and Dimension-Corrected Gamma in Educational Measurement Settings

Jari Metsämuuronen*
Finnish Education Evaluation Centre,
FINLAND

Received: October 23, 2020 ▪ Revised: December 10, 2020 ▪ Accepted: February 9, 2021

Abstract: Although Goodman–Kruskal gamma (G) is used relatively rarely it has promising potential as a coefficient of association in educational settings. Characteristics of G are studied in three sub-studies related to educational measurement settings. G appears to be unexpectedly appealing as an estimator of association between an item and a score because it strictly indicates the probability to get a correct answer in the test item given the score, and it accurately produces perfect latent association irrespective of distributions, degrees of freedom, number of tied pairs and tied values in the variables, or the difficulty levels in the items. However, it underestimates the association in an obvious manner when the number of categories in the item is more than four. Towards this, a dimension-corrected G (G2) is proposed and its characteristics are studied. Both G and G2 appear to be promising alternatives in measurement modelling settings, G with binary items and G2 with binary, polytomous and mixed datasets.

Keywords: *Item analysis, Goodman–Kruskal gamma, Somers D, Jonckheere–Terpstra test, Pearson correlation.*

To cite this article: Metsämuuronen, J. (2021). Goodman–Kruskal Gamma and dimension-corrected Gamma in educational measurement settings. *International Journal of Educational Methodology*, *7*(1), 95-118. https://doi.org/10.12973/ijem.7.1.95

Introduction

The family of gamma, delta, and tau

Two approaches are mainly used in estimating the association between two variables: one based on probability that has a latent linear nature and one based on covariance that has a latent trigonometric nature. In the former approach, most commonly used measures of association come from the family that include Kendall's tau-a and tau-b (Kendall, 1938, 1948), Goodman–Kruskal gamma (*G*; Goodman & Kruskal, 1954), and Somers' delta (*D*; Somers, 1962). In general, these coefficients estimate the probability that two randomly chosen respondents are in the same order in two variables (e.g., Van der Ark & Van Aert, 2015), or, in item analysis, they indicate what is the probability to obtain a correct answer in an item given a known score. The family of *G* and *D* also includes two other related measures, Kim's *d*y.x (Kim, 1971) and Wilson's *e* (1974) that are used, although rarely, for the same purpose as *G* and *D* (see Gonzalez & Nelson, 1996). Similarities of these coefficients are discussed later.

G and *D* are used relatively less often than the ones based on covariance, for example, product–moment correlation coefficient (PMC; Bravais, 1844; Galton, 1989; Pearson, 1896), often called Pearson correlation or sometimes Bravais–Pearson correlation (e.g., Cleff, 2017). However, this family of coefficients can be taken as the general case of many classic test statistics or those that are transformations of tau, *G* or *D* (see Newson, 2006). These include, among others, the sign test (Arbuthnott, 1710, see Conover, 1980; Metsämuuronen, 2017), the Gini index (Gini, 1912), the area under receiver operating characteristic (ROC) curve (AUC; e.g., Harrell, 2001; Heagerty & Zheng, 2005), Gehan–Breslow test (Breslow, 1970; Gehan, 1965) based on Wilcoxon W-test (Wilcoxon, 1945) and Kruskal–Wallis test (Kruskal & Wallis, 1956), Harrell's *C* index (Harrell et al., 1982), and Mann–Kendall trend test (Kendall, 1948; Mann, 1945). Also, the most popular nonparametric technique for estimating a linear trend (as described by El-Shaarawi & Piegorsch, 2001), Theil–Sen estimator (Sen, 1963; Theil, 1950) also known as Theil median slope and Kendall–Theil robust line, is defined through Kendall's tau-a, which can be thought of as a general case of *G* and *D*. Hence, the family of tau, *G* and *D* is an

* **Correspondence:**
Jari Metsämuuronen, Finnish Education Evaluation Centre, P.O. Box 28 (Mannerheiminaukio 1 A, 6th Floor), FI-00101 Helsinki, Finland.
✉ jari.metsamuuronen@gmail.com

interesting one from the viewpoint of multiple applications with a variety of options.

G and D in educational settings

In general, when two variables are measured on the ordinal or interval scale, *G* is a suitable measure of association of the variables. In educational settings, *G* is versatile when it comes to applications. *G* has been used, as examples, when analyzing the association between educational background and attitudes toward education or socioeconomic status and intelligence (Metsämuuronen, 2017), when analyzing college students' reactions to instruction and courses that use educational technology (Good, 2015), and association between sex, number of siblings, and participation in higher education (Shafina, 2021) just to mention a few. Higham et al. (2016) studied internal mapping and metacognitive accuracy and noted that *G* is, by far, the most used coefficient of association in these settings. Discussion section presents some further advances of *G* and dimension-corrected *G* derived later in the article.

In educational measurement settings, *G* and partial *G* (Goodman & Kruskal, 1954; see also Davis, 1967) are used, to some extent although rarely, in item analysis (e.g., Forthmann et al., 2020; Kreiner & Christensen, 2009; Nielsen et al., 2017; Nielsen & Santiago, 2020). However, remembering that the coefficients *G* and *D* are close siblings, they have a strict connection to educational measurement settings because of their connection to rank-biserial correlation coefficient (Cureton, 1956) based on *U* test statistic (Mann & Whitney, 1947) and rank-polyserial correlation (Metsämuuronen, 2021) based on Jonckheere–Terpstra test statistic (*JT;* Jonckheere, 1954; Terpstra, 1952); these are special cases of *D* and *G* (Newson, 2008; Metsämuuronen, 2021). Metsämuuronen (2020b) based on Newson (2002) and Metsämuuronen (2020a) reminds us that *D* correctly reaches the values +1 and −1, it is stable with extreme values, and it gives estimates for item discrimination power (IDP, see Lord & Novick, 1968) that are remarkably closer the real association between item and score than the widely used estimators item−total correlation (*Rit* = PMC) and item−rest correlation (*Rir*, Henrysson, 1963). In the practical educational testing settings (see, e.g., Aslan & Aybek, 2020; Delil & Ozcan, 2019), when the sample sizes may be small and the normality in the score cannot be ensured, these are advances.

Because *D* and *G* are close siblings, the positive characteristics of *D* are expected to also be incorporated in *G*. Then, whenever *D* is used, *G* could have been used also. For example, even though in Metsämuuronen and Ukkola (2019) reliability of extremely easy and difficult tests of 1st graders achievement levels was analyzed by replacing *Rit* with a less-underestimating option *D* in the formula of coefficient alpha, *G* could have been used also. The latter option is discussed in Discussion. However, the characteristics of *G* in educational measurement settings are mainly unstudied as are its capability to reach the true association between item and score and its potential character of underestimate IDP with polytomous items that is characteristic to Somers' *D* (e.g., Metsämuuronen, 2020b). Hence, the characteristics of *G* are worth studying in measurement modeling settings. This article aims to cover these gaps.

Research questions

The characteristics of *G* are studied within educational measurement modelling settings in three sub-studies. Study 1 is about the capability of *G* to reflect the true association between two variables under the specific condition in measurement modelling settings that a common latent trait drives both the score and the item. Study 2 is about with the underestimation of association in *G* from the theoretical viewpoint by connecting *G* with Greiner's relation on the one hand and by using empirical datasets with real-world items on the other. Studies 1 and 2 show that although *G* is accurate in reflecting the latent probability that the pairs of test takers are in the same order in the item as they are in the score, it underestimates the association in an obvious manner when the number of categories in an item are more than four. Hence, in Study 3, to enhance *G* for the polytomous datasets, a dimension-corrected *G* (G_2) is derived and its characters are studied in relation to the relevant estimators *G*, *D* and D_2, PMC, and polychoric correlation coefficient (R_{PC}). Next section reviews the known characteristics of *G* relevant here.

Some known characteristics of G

Sample forms of G and D

In general, both *G* and *D* estimate the probability that two randomly chosen cases have the same order in two variables (γ and δ, respectively; e.g., Van der Ark & Van Aert, 2015; Metsämuuronen, 2021). In measurement modelling settings, this is sometimes interpreted as the relationship between test score and the probability to choose the correct response (Forthmann, et al., 2020 based on Love, 1997).

Let (x_1, y_1), …, (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, $x_l < x_h$ and $y_l < y_h$ or $x_l > x_h$ and $y_l > y_h$. The pairs are discordant when $x_l < x_h$ and $y_l > y_h$ or $x_l > x_h$ and $y_l < y_h$. If $x_l = x_h$ or $y_l = y_h$, the pairs are tied, that is, they are neither discordant nor concordant.

Let item *g* and the score *X* form an $R \times C$ cross-table with cell frequencies n_{ij}. We define

$$C_{ij} = \sum_{h<i}\sum_{k<j} n_{hk} + \sum_{h>i}\sum_{k>j} n_{hk},$$

$$D_{ij} = \sum_{h<i}\sum_{k>j} n_{hk} + \sum_{h>i}\sum_{k<j} n_{hk},$$

$$P = \sum_{i,j} n_{ij} C_{ij},$$

and

$$Q = \sum_{i,j} n_{ij} D_{ij} \quad (1)$$

In estimating γ and δ, what is crucial are the number of pairs in the same direction (P), and the number of pairs in the opposite directions (Q). Notably, the quantities of P and Q by Eq. (1) are double of those we usually see in textbooks (e.g., Metsämuuronen, 2017; Siegel & Castellan, 1988). This notation above leads us to the correct asymptotic standard errors (e.g., Agresti, 2010; Goodman & Kruskal, 1979; see Study 3) instead of rough approximation (e.g., Somers, 1980; Metsämuuronen, 2017; Siegel & Castellan, 1988).

As a benchmark of G, three outcomes are expected in calculating D, based on whether g or X is dependent, or we are interested in the symmetric association. In measurement modelling settings, the direction "g given X" is relevant because the testing theory postulates that the latent trait manifested as the score explains the behaviour in the item and not vice versa (e.g., Byrne, 2016; Metsämuuronen, 2017).† The number of all combinations of pairs related to this direction is

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

where T_g denotes the tied pairs common to both P and Q. Gonzalez and Nelson (1996) connect these ties with the predictor variable (T_p) although those seem to be related to the widths the scales rather than to predictor or criterion status of the variable (see Metsämuuronen, 2021).

G proportions $P - Q$ with *relevant* pairs, that is, with only those pairs where we *know* the direction:

$$G = \frac{P-Q}{P+Q} = \frac{P-Q}{D_r - 2T_g}. \quad (3)$$

Notably, while calculating G, the tied pairs are excluded in the same manner as done in both sign test and Wilcoxon signed-rank test.

D proportions $P - Q$ with *maximal* number of pairs to the same direction including the number of tied pairs:

$$D\left(g \middle| X\right) = D = \frac{P-Q}{D_r} = \frac{P-Q}{P+Q+2T_g}. \quad (4)$$

Kim's (1971) $d_{X.g}$ equals this form of Somers' D. Henceforth, this form of Somers' D is called D.

Usually, the datasets include tied pairs between the variables. Then, $P + Q < P + Q + 2T_g$ and the magnitude of the estimates by G is higher than those by D; G gives us a more liberal estimate while D gives us more conservative estimate of the probability that two cases are in the same order in g and X.

Connection of G and D with Jonckheere–Terpstra test statistic

From the interpretation viewpoint of both G and D, their connection with the Jonckheere–Terpstra test statistic (JT) is worth noting. Metsämuuronen (2021) showed that

$$D = 2 \times \frac{JT}{\sum_{i<j} n_i n_j} - 1 \quad (5)$$

† In the literature related to directional coefficients (e.g., IBM, 2017; Newson, 2002, 2006; Siegel & Castellan, 1988), this direction is called "X dependent" and it is notated as ($X|g$). This logic comes from the GLM settings with eta squared where the metric variable (X) such as achievement cannot explain the nominal variable, e.g., the gender (g) and, hence, X is "dependent" and, consequently, g must be "independent". However, within the measurement modelling settings, the same direction means that the latent trait manifested as the score (X) explains the *order* in the item (g). In these settings the relation of g and X is thought from conditional viewpoint as "g given X" ($g|X$). In this article, this logic familiar from the conditions is used in the notation: $D(g|X)$ refers to "delta so directed that 'g given X'" which, in the outputs of some generally known software packages such as IBM SPSS, SAS, as well as the *R* libraries, would be labelled "X dependent".

and

$$G = 2 \times \frac{JT}{\sum_{i<j}^{g} n_i n_j - T_g} - 1' \qquad (6)$$

where $\sum_{l<h}^{g} n_l n_h$ refers to the maximal number of pairs in one direction. Because *JT* indicates the *number* of logically ordered cases in *g* after they are ordered by *X*, *D* and *G* indicate a slightly modified *proportion* of logically ordered cases in *g* after they are ordered by *X*.

Directional nature of G

Somers' *D* takes three values depending on whether the row or column is dependent or whether the association is symmetric. On the contrary, *G* provides us only one estimate and, hence, traditionally, it has been taken as a symmetric measure (e.g., IBM, 2017; Sheskin, 2111; Sirkin, 2006; Wholey et al., 2015). However, Metsämuuronen (2021) showed that, under certain conditions, $G = D(g|X)$ and no other way. Also, if $T_g, T_X > 0$, *G* follows closer to $D(g|X)$ than to $D(X|g)$ although always $G > D(g|X)$. That is, *G* is unambiguously a directional measure in the same direction as D("*g* given *X*") as in conditions or *D* ("*X* dependent") as in GLM settings. This direction makes sense in measurement modelling settings.

Kvålseth (2017, p. 10582; see also Higham & Higham, 2019; Masson & Rotello, 2009) 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". Several solutions have been proposed to correct *G*, (e.g., Bai & Wei, 2009; Highan & Higham, 2019; Hryniewicz, 2006; Kvålseth, 2017; Masson & Rotello, 2009; Rousson, 2007). However, Freeman (1986), Gonzalez and Nelson (1996), and Metsämuuronen (2021) propose that there is no "inflation" per se in *G; G* accurately reports a slightly modified proportion of logically ordered test-takers in the item after they are ordered by the score, taking into account only the pairs where the direction is known. The apparent "inflation" may be caused by the hidden directional nature of *G*.

PMC as a benchmark to G and D in item analysis settings

In item analysis settings, PMC = *Rit*, being one of the traditional indicators of IDP, has the interesting characteristic to *always* underestimate association between an item and score, and because of this, it can be used as a specific benchmark of an *obvious* underestimation of association for the other estimators.

The reason for the underestimation of association by *Rit* is the phenomenon called the restriction of range (RR) related to PMC (see the literature in Meade, 2010; Sackett et al., 2007; Sackett & Yang, 2000). It is a known characteristic of PMC, already discussed by Pearson (1903), that when only a portion of the range of values of a (latent) variable is actualized in the sample, it affects the inaccuracy of the true association (see simulations in Martin, 1973, 1978; Olson, 1980). In practical terms, when the scales of two variables differ from each other, as is common in measurement modelling settings with an item and a score, PMC *always* underestimates the true association between *g* and *X* (see algebraic reasons in, e.g., Metsämuuronen, 2016, 2017). Then, the magnitude of the estimates that are lower than those by PMC is indicative of an *obvious underestimation* of the association.

Although *D*, being appealing in item analysis settings because it underestimates the IDP remarkably less than PMC does with binary items (Metsämuuronen, 2020a), it tends to underestimate IDP more than PMC when the item has three categories or more (Metsämuuronen, 2020b). Therefore, Metsämuuronen (2020b) proposed a dimension–corrected *D* (D_2) that gives estimates that underestimate IDP less than PMC and R_{PC} do without giving obvious overestimates. Because of the close relation of *D* and *G*, it is expected that *G* also underestimates IDP when the number of categories in the item increase. This is examined in Study 2.

Study 1: Capability of G to reflect the true association between g and X

Research question in Study 1

Study 1 examines the extent to which *G* reflects the true association between two variables under the condition that is specific to measurement modelling settings that a common latent variable θ drives both item and score causing the true association between the item and the score to be always perfect. The behaviour of *G* is compared with *D*, PMC, and R_{PC} by varying the latent variables (normal, skewed normal, and even), $df(X) = C - 1$, $df(g) = R - 1$, and the difficulty level of *g* (*p*).

Measurement and statistical models related to gamma

The models based on latent variable modeling (see Raykov & Marcoulides, 2013) assume that a latent trait or a latent variable θ drives the observed responses in a test item g_i (x_i). Assume a widely used simplified‡ one-factor measurement model where θ, manifested as an observed variable of scores X (y_i) with C distinctive ordinal or interval categories, explains the behaviour in g_i with R distinctive ordinal or interval categories and where $R << C$. θ is linked with x_i by the weighting factor w ($-1 \le w_i \le +1$) and the error related to the item (e_i): $x_i = w_i\theta + e_i$ (e.g., Cheng et al., 2012; McDonalds, 1985). The linking element is usually a coefficient of association, and, traditionally, has been interpreted as factor loading in factor analysis and structural equation modelling settings. However, in general, the weight element could be any coefficient of association such as PMC, R_{PC}, or G.

From the statistical model viewpoint, assume that the observed values in g_i with x_i = 1, …, R and X with y_i = 1, …, C distinctive ordinal or interval categories, and $R << C$, share the common latent trait (θ). Then, the higher the latent trait is the more probable it is to reach a higher score (X) and, simultaneously, more probably the correct answer (or a higher value) in the test item. The threshold values of θ for each category in g_i are denoted by υ_i and for each category in X by τ_j: $g = x_i$, if $\upsilon_{i-1} \le \theta < \upsilon_i$, i = 1, 2,…, R and the observed value of the score $X = y_j$, if $\tau_{j-1} \le \theta < \tau_j$, j = 1, 2, …, C, and $\upsilon_0 = \tau_0 = -\infty$ and $\upsilon_R = \tau_C = +\infty$. Figure 1 illustrates the statistical model with a binary g (R =2); n_{ij} refers to the number of cases in in cell i, j.

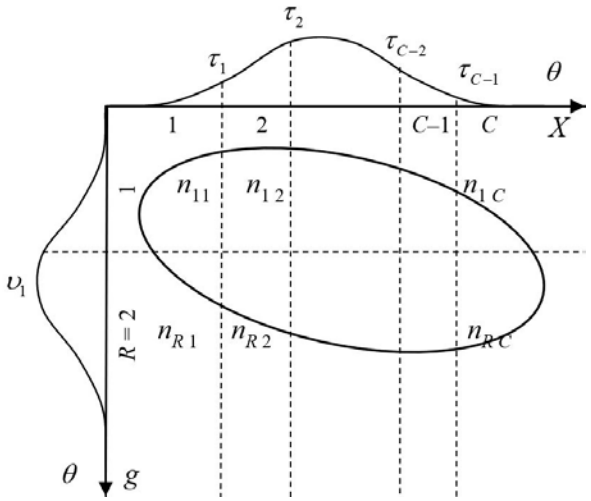


Figure 1. A latent variable θ manifested in two ordinal variables g and X

The theoretical condition of simplified one latent common trait causes the true association between g and θ to be perfect.§ The task in Study 1 is to examine how well G—as well as the related benchmarking coefficients—can detect this latent perfect association.

Dataset used in Study 1

For the simulation, three vectors with N = 1000 cases were created: a normal vector with N(0,1), a skewed-normal vector with Γ(2,1), and an even vector. Each vector was duplicated to form a pair of perfectly correlated variables. These pairs of vectors were manipulated so that one became a variable with a narrower scale (item g) and the other with a wider scale (score X). The scale of X related to normal and gamma distributions was set to vary with the fixed values $df(X)$ = 4, 6, 12, 20, 25, 30, 40, and 60 and the even distribution with $df(X)$ = 4, 9, 19, 24, 39, 49, and 99. The scale of g was set to vary with fixed values $df(g)$ = 1, 2, 3, and 4, that is, the most commonly used scales from binary to 5-point Likert type of scales were covered.

Results: G reflects the perfect true perfect association without loss of information

Figures in Appendix illustrate the difference between the estimators if $df(g)$ = 1; the graphs with $df(g)$ = 2–4 would give, essentially, identical information of the relation. Notably, in all conditions, G and R_{PC} reproduce the perfect association between the item and the score while the estimates by D and Rit either underestimate the true association

‡ Obviously, *several* independent (or dependent) latent factors such as general intelligence, attitude toward the test, perseverance, and reading ability are related to the item responses in real-life settings. Hence, the common one-factor model is highly theoretical condition.

§ Notably though, in real-life testing settings (see Study 2), the association, in fact, is rarely deterministic and we do not expect to see perfect association.

or behave unpredictably, specifically, with short tests. The reason why *D* and PMC behave unpredictably with short tests is related to the issue of tied pairs: *G* ignores tied pairs while *D* uses them, and PMC also counts the covariance in such cases.

Let us take an item with $p = 0.5$, $df(g) = 1$, $df(X) = 6$ with the latent normality as an example (Table 1). Given the cross-table in Table 1, PMC = 0.728, $D = 0.841$, $G = 1.000$, and $R_{PC} \approx 1.000$.[**]

Table 1. A pair of variables with perfect latent association forming 2 × 7 Crosstable

		X							**Total**
Count		**0**	**1**	**2**	**3**	**4**	**5**	**6**	
g	0	5	54	242	199	0	0	0	500
	1	0	0	0	199	242	54	5	500
Total		5	54	242	398	242	54	5	1000

In practical educational testing settings, if one of the general statistical software packages is in use, *G* and *D* are simple to calculate. In IBM SPSS, the syntax is CROSSTABS /TABLES=item BY Score /STATISTICS=GAMMA D. In SAS, the command PROC FREQ provides *G* and *D* by specifying the TEST statement by GAMMA, SMDCR options. Correspondingly, after defining *C* and *D*, RStudio, as an example, uses the syntax SomersDelta (x, y = NULL, direction = c("row", "column"), conf.level = NA, ...) for D and calc.gamma <- function(x), { x <- matrix(as.numeric(x), dim(x)), c <- concordant(x), d <- discordant(x), gamma <- (c - d) / (c + d) for *G* (see, e.g., https://gist.github.com/marcschwartz/3665743).

The manual calculation can be done as follows. For *G* and *D, P* and *Q* are calculated the same manner, that is, given Table 1,

$$\begin{aligned} P &= 2\times\big((5+54+242)\times(199+242+54+5)+199\times(242+54+5)\big) \\ &= 2\times 210,399 \\ &= 420,798 \end{aligned}$$

and

$Q = 0$.

G ignores the tied pairs of which the direction is not known and, hence, $G=(P-Q)/(P+Q)=P/P=1.000$. The number of all pairs is $D_r = N^2 - \sum_i n_i^2 = 1000^2 - 2\times 500^2 = 500,000$. Then, $D=(P-Q)/D_r$ = 420,798/500,000 = 0.841.

Study 2: Underestimation of the association by G with the polytomous items

Research question in Study 2

The second research question is: under which conditions *G* underestimates IDP. The matter is first considered from the theoretical viewpoint by connecting *G* with Greiner's relation. Second, empirical dataset is used to study the phenomenon with real-world items.

Underestimation by G from the theoretical viewpoint

Obvious underestimation of IDP by *G* is expected because of Greiner's relation (Greiner, 1909) discussed by Kendall (1949), Newson (2002), and Metsämuuronen (2020b). With continuous variables, *G* = *tau-a* and, then, Greiner's relation states that

$$G = Tau_a = \frac{2}{\pi}\arcsin\left(\rho_{gX}\right). \qquad (7)$$

Relation of PMC and *G* is illustrated in Figure 2. We note the linear nature of *G* in relation to the trigonometric nature of PMC.

[**] To estimate R_{PC}, two alterations are needed in the traditional procedure (see, Drasgow, 1986). First, PMC embedded in the process cannot take the actual value 1 although a value as close to 1 as possible, such as $0.\overline{99}$, can be allowed. Second, as small positive number as possible, such as 10^{-50}, should be added to each logarithm term because logarithm cannot be taken of a zero. Hence, technically, R_{PC} cannot reach the value 1 but it can be very close. For the analysis of R_{PC} in this article, Zaiontz' (2021) procedure for two-step estimation for manual calculation was used.

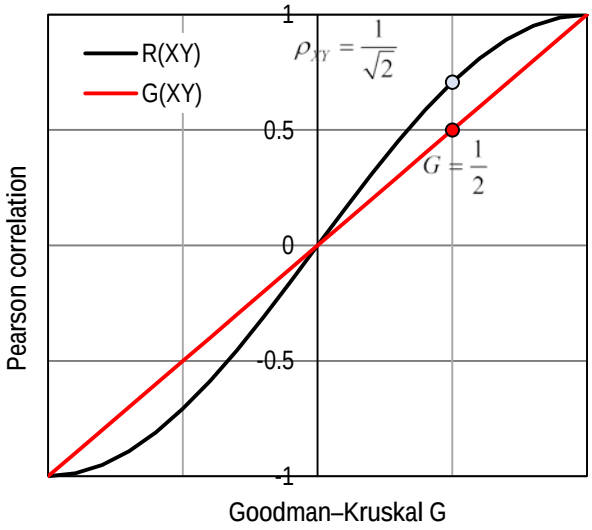


Figure 2. Relation of PMC (R_{XY}*) and Goodman–Kruskal* G

From Eq. (7), it is known that, in the case of two continuous variables, except for the extreme values ±1 and 0, the magnitude of the estimates by PMC tends to be greater than those by *G*; for *G* = 0.5 we would expect to see PMC = $1/\sqrt{2} = 0.7071$ (see Fig. 2). Hence, we may predict that, if the association is not perfect (or near perfect) and when the number of categories increase, *G* tends to underestimate the true association more than PMC.

Datasets used in Study 2

In Study 2, the connection of *G* and PMC is studied by using two larger real-world datasets. Both are based on a national-level dataset of 4,022 test-takers of a mathematics test with 30 binary items (Finnish National Education Evaluation Centre [FINEEC], 2018). In this original dataset, the lower bound of reliability was α = 0.885, item discrimination estimated by PMC = *Rit* ranged $0.332 < \rho_{gX} < 0.627$ with the average $\overline{\rho_{gX}} = 0.481$, and the difficulty levels of the items ranged 0.24 < *p* < 0.95 with the average $\overline{p}$ = 0.63.

For the training dataset, ten random samples of *n* = 50, 100, and 200 test-takers were picked from the original dataset. In each of the 30 datasets, 36 shorter tests were produced by varying the number of items, difficulty levels of the items, and *df*(*g*) and *df*(*X*). Polytomous items were constructed as compilations of the original binary items. Thus, the training dataset consisted of 11,160 test items from 1,080 tests with a varying number of test-takers (*N* = 50, 100, and 200) and items (*k* = 2–30), difficulty levels ($\overline{p}$ = 0.55–0.76), reliabilities (α = 0.739–0.935), and degrees of freedom in the item (*df*(*g*) = 1–15), and in the score (*df*(*X*) = 12–27). As benchmarks of *G*, the estimates by PMC and *D* were produced for all items and R_{PC} for half of the datasets (*k* = 5,580). Selected characteristics of the items are collected in Table 2.

The training dataset was limited to relatively short tests (*df*(*X*) < 28). Hence, another dataset called "cross-validation dataset", partly artificial, was prepared. The same original dataset was used, however, such that 30 items from it were duplicated and the response patterns of a portion of real test-takers was changed to cause mild changes in item difficulties and item–total correlations. The modified items were combined with the authentic ones to form a dataset with 60 binary items as parallel tests with odd–even items. For this dataset, 19 sets of *n* = 200 test-takers were picked, and 72 subtests in each set with the number of items of *k* = 30, 35, 40, 45, 50, 55, and 60 were produced. Altogether 19×72 = 1,368 sub-tests ending up to *k* = 29,887 items were produced. In this dataset, *df*(*X*) = 18–42. In Study 2, this dataset is referred to if the results between the datasets differ radically from each other. Both datasets are used also in Study 3.

Results: G underestimates association in an obvious manner when there are more than 4 categories in an item

Comparing Eqs. (3) and (4), the magnitude of the estimates by *G* are generally higher than those by *D*. Because of Eq. (7), obvious underestimation of IDP by *G* is expected when the number of categories in the item increase. Then, a relevant question is, what is the threshold number of categories for *G* to underestimate IDP? In the training dataset, this threshold appears to be four or five categories (Table 2; Figure 3); when the item has five categories or more (*df*(*g*) ≥ 4), *G* tends to give obvious underestimates of IDP. In this regard, the cross-validating dataset suggest four categories.

Table 2. Selected characteristics of 11,160 items in the training dataset

	Average Association				Standard deviation				N	
***df*(*g*)**	**PMC**	***G***	***D***	R_{PC}	**PMC**	***G***	***D***	R_{PC}	**PMC, *G*, and *D***	R_{PC}
1	0.486	0.630	0.608	0.628	0.115	0.139	0.137	0.132	5943	2972
2	0.625	0.672	0.647	0.703	0.088	0.093	0.093	0.090	2265	1129
3	0.708	0.711	0.684	0.759	0.069	0.075	0.074	0.067	1029	514
4	0.771	0.748	0.718	0.807	0.056	0.060	0.060	0.051	546	272
5	0.812	0.770	0.739	0.835	0.050	0.057	0.056	0.051	354	183
6	0.846	0.794	0.763	0.864	0.032	0.041	0.040	0.033	278	137
7	0.873	0.817	0.784	0.887	0.029	0.038	0.037	0.029	190	95
8	0.892	0.834	0.800	0.902	0.025	0.034	0.034	0.028	98	52
9	0.911	0.856	0.822	0.919	0.026	0.040	0.038	0.023	127	63
10	0.926	0.872	0.837	0.933	0.020	0.030	0.028	0.023	118	58
11	0.940	0.888	0.854	0.944	0.013	0.021	0.023	0.011	85	42
12	0.943	0.887	0.854	0.945	0.010	0.018	0.020	0.009	61	33
13-14	0.944	0.886	0.856	0.947	0.008	0.014	0.015	0.006	66	30
Total	0.596	0.675	0.650	0.696	0.168	0.131	0.129	0.140	11,160	5,580

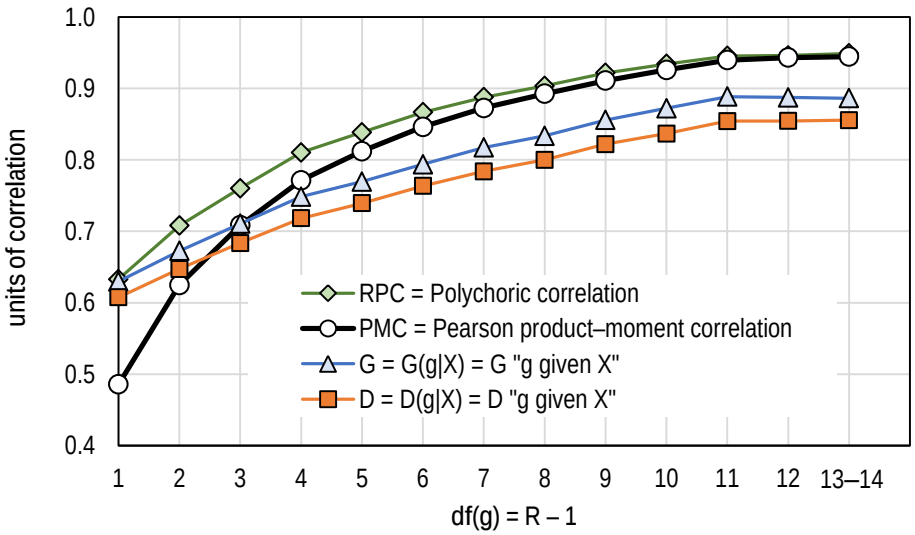


Figure 3. Average estimates by PMC, D, G *and* R_{PC} *by df(g) (k = 11,160 items)*

Three points of the relation of measures based on probability (*G* and *D*) and covariance (PMC and R_{PC}) are highlighted. First, with binary items, the magnitude of the estimates by *G* are markedly higher than those by PMC and they correspond quite closely with those by R_{PC}. Of the 5,943 estimates with *df*(*g*) = 1, only in four (0.1%), *G* < PMC. Knowing that the estimates by PMC always underestimate the true association and, if the estimates by R_{PC} do not *over*estimate the association, in the binary settings, the estimates by *G* tend to be remarkably closer to the true association than those by PMC.

Second, the higher the number of categories of the item gets the more probable it is to find PMC > *G*; with three categories, 3.3% of the items showed PMC > *G.* The parallel figure with four categories is 48.4%, and with five categories, 89.6%.

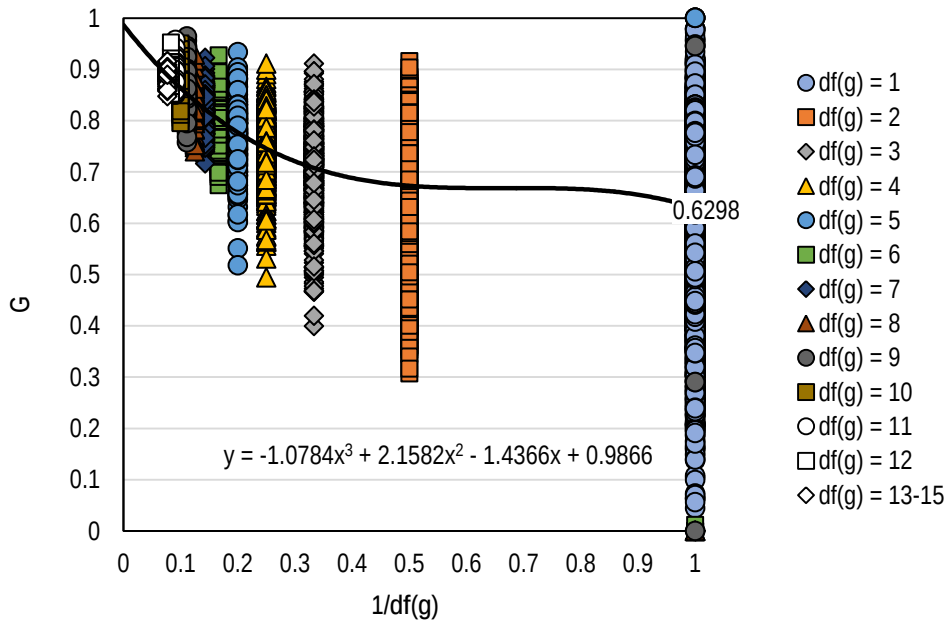


Figure 4. Relation of G *and 1/*df*(*g*)*

Third, the tendency of *G* to produce estimates with a magnitude lower than PMC when *df*(*g*) ≥ 4 seen in Figure 3 leads us to the illusion that, at this range, always, *G* < PMC. However, Figure 4 illustrates the practical specialty embedded in *G* and PMC as well as in all estimators of association in item analysis settings that the estimate approximates the perfect 1 the less there are items in the test and the more there are categories in the items. The average magnitude of the estimates by *G* can be modelled as a third-grade equation where 1/*df*(*g*) explains *G*. This model is elaborated on when deriving the dimension-corrected *G* in Study 3.

Study 3. Dimension-corrected gamma

Research question in Study 3

It was shown in Study 1 that, unlike *D* and PMC, *G* is accurate in reproducing the perfect latent true association between the item and the score. Study 2 showed that, when the number of categories in the item exceeds four, the estimates by *G* tend to underestimate the IDP in an obviously manner. Hence, although *G* is accurate in reflecting the *probability* that the test-takers are in the same order in both the item and the score, it seems that the probability of same order as an indicator of IDP leads to obvious underestimates in comparison with covariance between the variables. Hence, it makes sense to develop a "dimension-corrected *G*" that would turn the linear nature of *G* into more trigonometric. If it behaves same as D_2, this transformation would overcome the disadvantage of obvious underestimation by *G*.

In what follows, first, some basic underlying principles for the derivations are discussed. Second, the underestimation in *G* is modelled based on the training dataset. Third, a dimension–corrected *G* (G_2), specific to measurement modelling settings, is suggested. Also, a corrected form for D_2 is suggested based on G_2. Fourth, the characteristics of G_2 are studied in relation to *G*, *D*, D_2, PMC, and R_{PC}.

Principles underlying the modelling of the dimension-corrected G

Based on Study 1 and 2, underlying the process of deriving the correction elements, six main notes (N) are made and five consecutive principles (P) are followed:

N1. PMC always underestimates IDP in item analysis settings when $df(g) << df(X)$ (Study 1).

P1. The estimate by G_2 should be higher than that by PMC to overcome the obvious underestimation by *G*.

N2. *G* reflects accurately the true association under the assumptions related to measurement modelling settings (Study 1).

P2. With the deterministic patterns between *g* and *X*, *G* need not to be corrected.

N3. *G* gives a credible estimate of IDP when *df*(*g*) = 1 (Study 2).

P3. *G* should be corrected only when *df*(*g*) > 1.

N4. *G* tends to underestimate IDP the higher is the *df*(*g*) (Study 2).

P4. Dimension-correction in G_2 should affect more correction the higher the *df*(*g*) is.

N5. In real-life settings, *G* reaches the maximal value 1 while PMC does not (Studies 1 and 2).

P5. When *G* = 1, no correction is needed. Additionally, obviously, G_2 should not exceed 1.

N6. *G* correctly reaches the value 0.

P6. When *G* = 0, no correction is needed.

Modelling the underestimation in G

The dimension-corrected *G*, henceforth G_2, is based on modelling the underestimation in 11,160 empirical values of *G* in the training dataset. Figure 5 illustrates the starting point of the modelling.

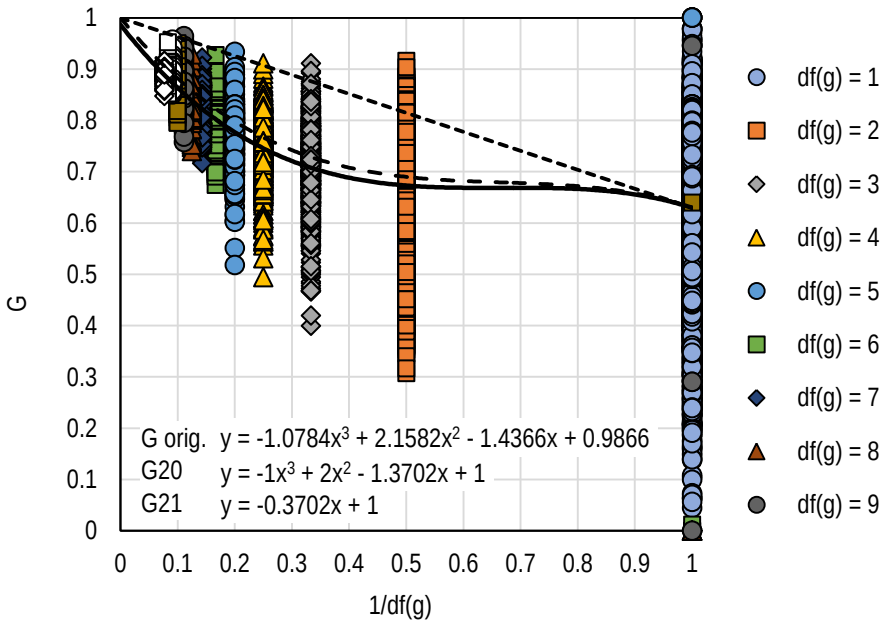


Figure 5. The original model of G *and initial models* G_{20} *and* G_{21}

The training dataset suggests that the model with cubic nature $-1.08/df(g)^3 + 2.16/df(g)^2 - 1.44/df(g) + 0.99$ has the best fit between the observed distribution of *G* and 1/*df*(*g*). However, this model is somewhat misleading because the polynomial curve should go through the points (1/*df*(*g*) = 0, *G* = 1) and (1/*df*(*g*) = 1, *G* = 0.62979). The first point indicates that, when there is only one item in the test, this item correlates perfectly with the "score" formed by this item causing *G* = 1. The second point refers to the expectation of the level *G* when *df*(*g*) = 1. During the derivation, the latter value is diminished. Hence, the final correction does not depend on the factual average value of *G* in *df*(*g*) = 1.

Notably, the original model in the cross-validating dataset differs from the one obtained in the training dataset to a certain extent. The average estimate of *G* when *df*(*g*) = 1 in the cross-validating dataset (0.461) is significantly and notably smaller in magnitude than in the training dataset (0.630). The differences between the models leads to the realization that G_2 derived in what follows seems to give conservative in the correction when the scale of the score exceeds 30 categories.

The corrected model G_{20} passing through the points (1/*df*(*g*) = 0, *G* = 1) and (1/*df*(*g*) = 1, *G* = 0.62979) is:

$$
\begin{aligned}
G_{20} &= 1 - \frac{1.37021}{df(g)} + \frac{2}{df(g)^2} - \frac{1}{df(g)^3} \\
&= 1 - \frac{0.37021}{df(g)} - \left(\frac{1}{df(g)} - \frac{2}{df(g)^2} + \frac{1}{df(g)^3} \right) \\
&= 1 - \frac{0.37021}{df(g)} - \frac{1}{df(g)} \left(1 - \frac{1}{df(g)} \right)^2 ,
\end{aligned}
\tag{8}
$$

where 0.37021 = 1 – 0.62979. The magnitude of the underestimation is unknown. For modelling purposes, the "correct" level of G (G_{21}) was set to be linear through the points ($1/df(g) = 0$, $G = 1$) and ($1/df(g) = 1$, $G = 0.62979$:

$$G_{21} = 1 - \frac{0.37021}{df(g)}. \tag{9}$$

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

$$\begin{aligned} G_E &= G_{21} - G_{20} \\ &= 1 - \frac{0.37021}{df(g)} - \left(1 - \frac{0.37021}{df(g)} - \frac{1}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2\right) \\ &= \frac{1}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2 \end{aligned} \tag{10}$$

and, hence, the initial correction for G is

$$G_{22} = G + G_E = G + \frac{1}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2. \tag{11}$$

In the second phase, three switches are added to G_E: $\left(df(g) - 1\right)$ related to P3 to restrict the correction only on the items with $df(g) > 1$, $(1 - G)$ related to the principle P5 to restrict the correction only on items with non-deterministic patterns, and $(G - 0)$ related to the principle P6 to restrict the correction only on items with non-zero association. After these, a suggested correction factor is

$$\left(df(g) - 1\right) \times (1 - G) \times (G - 0) \times G_E = \left(df(g) - 1\right) \times \left(G - G^2\right) \times G_E. \tag{12}$$

Then, combining Eqs. (10) and (11), a suggestion as the dimension-corrected G is

$$\begin{aligned} G_2 &= G + \left(G - G^2\right) \times \frac{\left(df(g) - 1\right)}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2 \\ &= G \times \left(1 + (1 - G) \times \frac{\left(df(g) - 1\right)}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2\right) \end{aligned} \tag{13}$$

Eq. (13) can be further modified into form

$$G_2 = G \times \left(1 + (1 - G) \times A\right) \tag{14}$$

where

$$A = \frac{df(g) - 1}{df(g)}\left(1 - \frac{1}{df(g)}\right)^2. \tag{15}$$

The correction in Eq. (13) fits the positive values of G. Because of the symmetry in the values of G, a more general form of G_2, comprising also the negative values of G, is

$$G_2 = G \times \left(1 + \left(1 - abs(G)\right) \times A\right). \tag{16}$$

Notably, the key element A for the dimension correction in Eq. (14) is the same as in D_2 (Metsämuuronen, 2020b). However, G_2 includes one switch more than D_2 (the element $G - 0$) and, hence, if $D = 0$, the form of D_2 presented in Metsämuuronen (2020b) leads automatically to a non-zero estimate when $df(g) > 1$. Therefore, it would be better to use the same correction elements as in G_2 also in D_2. Then, a corrected form of D_2 parallel to G_2 is:

$$D_2 = D \times \left(1 + \left(1 - abs(D)\right) \times A\right). \tag{17}$$

This corrected version of D_2 is used in the comparison of the estimates in what follows. Because of Eqs. (3) and (4), except in the case when there are no tied pairs, the magnitude of the estimates by G_2 exceed those by D_2.

Limits of G_2

When $df(g) = 1$, $G = -1, 0, +1$, $G_2 = G$, otherwise $G_2 > G$. In the theoretical extreme case that $df(g) = \infty$, that is, with a continuous item and infinite number of test-takers with a unique item category (to form an infinite number of categories in the item),

$$\lim A = \lim \frac{df(g)-1}{df(g)}\left(1-\frac{1}{df(g)}\right)^2 = 1\times 1^2 = 1 \tag{18}$$

and, then, the correction in Eq. (15) seems to lead us to the triviality that, with continuous variables, $G_2 = 1$ seemingly irrespective of the actual association between the item and the score. However, with indefinitely long "parallel tests", the association between the sub-tests and the score approximates the ultimate magnitude of PMC = G_2 = 1. Hence, in item analysis settings, with indefinitely many categories in the item(s), the score would also contain indefinite number of categories and, then, G approximates the magnitude of 1. Nevertheless, Eq. (18) hints that when two variables with different scales are *independent* from each other, another kind of correction than provided by Eqs. (14) and (16) may be needed. *This limitation of* G_2 *is necessary to keep in mind if applying it to independent items.*

Asymptotic sampling variance and standard error of G2 and corrected D2

Because the statistical properties of G are well documented (e.g., Agresti, 2010; Goodman & Kruskal, 1979; Siegel & Castellan, 1988), the sampling variance and, hence, the asymptotic standard errors (ASE) of G_2 are known in the cases of $df(g) = 1$ and $G = \pm 1$ and 0 because, in these cases, $G_2 = G$:

$$\sigma^2_{G_2} = \sigma^2_G = \frac{16}{(P-Q)^4}\sum_{i,j} n_{ij}\left(QC_{ij} - PD_{ij}\right)^2 \tag{19}$$

that leads to asymptotic standard error

$$ASE_1(G_2) = ASE_1(G) = \frac{4}{(P+Q)^2}\sqrt{\sum_{i,j} n_{ij}\left(QC_{ij} - PD_{ij}\right)^2}\,. \tag{20}$$

and, under the hypotheses of independence,

$$ASE_0(G_2) = ASE_0(G) = \frac{2}{(P+Q)}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2} \tag{21}$$

where P, Q, C_{ij}, and D_{ij} are as defined in Eq. (1). Of these, the former is used when calculating confidence intervals and the latter when testing hypotheses. To derive the corresponding sampling variance for the case of $df(g) > 1$, by using Eqs. (13) and (14) and the basic laws of variance, we get

$$\begin{aligned}\sigma^2_{G2} &= VAR\left(G + \left(G - G^2\right)\times A\right) = VAR(G) + A^2 \times \left(VAR\left(G\right) - VAR\left(G^2\right)\right)\\ &= \left(A^2+1\right)\times VAR\left(G\right) - A^2 \times VAR\left(G^2\right)\end{aligned}\,. \tag{22}$$

where A is as in Eq. (15). With the range $0 \le X \le +1$, that is the normal range to use G_2, $VAR\left(X^2\right) \ge \left[VAR\left(X\right)\right]^2$. Then, we can get the higher boundary of the sampling variance:

$$\sigma^2_{G2} \le \left(A^2+1\right)\times VAR\left(G\right) - A^2 \times \left[VAR\left(G\right)\right]^2 = VAR\left(G\right)\times\left[1 + A^2 \times \left(1 - VAR\left(G\right)\right)\right]. \tag{23}$$

Consequently,

$$\begin{aligned}ASE_1(G_2) &\le \sqrt{VAR\left(G\right)\times\left[1 + A^2 \times \left(1 - VAR\left(G\right)\right)\right]}\\ &= \frac{4}{(P+Q)^2}\sqrt{\sum_{i,j} n_{ij}\left(QC_{ij} - PD_{ij}\right)^2} \times \sqrt{1 + A^2 \times \left(1 - \frac{16}{(P+Q)^4}\times\sum_{i,j} n_{ij}\left(QC_{ij} - PD_{ij}\right)^2\right)}\end{aligned} \tag{24}$$

and, under the hypotheses of independent variables,

$$\begin{aligned}ASE_0(G_2) \le \frac{2}{(P+Q)}&\sqrt{\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2} \times\\ &\sqrt{1 + A^2 \times \left(1 - \frac{4}{(P+Q)^2}\times\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}(P-Q)^2\right)}\end{aligned}\,. \tag{25}$$

Similar manner, the sampling variance of the corrected D_2 is

$$\sigma_{D2}^2 \le VAR(D)\times\left[1+A^2\times\left(1-VAR(D)\right)\right]. \tag{26}$$

In the dichotomous case, and when D = ±1 or 0, $D_2 = D$ and the asymptotic sampling variance of the corrected D_2 can be approximated as

$$\sigma_{D_2}^2 = \sigma_D^2 = \frac{4}{D_r^4}\sum_{i,j} n_{ij}\left(D_r\left(C_{ij}-D_{ij}\right)-\left(P-Q\right)\left(N-n_i\right)\right)^2 \tag{27}$$

which leads to asymptotic standard error usable when calculating the confidence intervals as

$$ASE(D_2 1) = ASE(D1) = \frac{2}{D_r^2}\sqrt{\sum_{i,j} n_{ij}\left(D_r\left(C_{ij}-D_{ij}\right)-\left(P-Q\right)\left(N-n_i\right)\right)^2} \tag{28}$$

and, under the hypotheses of independent variables, usable when testing hypotheses, as

$$ASE(D_2 0) = ASE(D0) = \frac{2}{D_r}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij}-D_{ij}\right)^2-\frac{1}{N}\left(P-Q\right)^2} \tag{29}$$

as in Metsämuuronen (2020b). However, when $df(g) > 1$ and D differs from 0 and 1, asymptotic standard errors are calculated parallel to those of G_2 as

$$\begin{aligned} ASE_1(D_2) &\le \sqrt{VAR(D)\times\left[1+A^2\times\left(1-VAR(D)\right)\right]} \\ &= \frac{2}{D_r^2}\sqrt{\sum_{i,j} n_{ij}\left(D_r\left(C_{ij}-D_{ij}\right)-\left(P-Q\right)\left(N-n_i\right)\right)^2}\times \\ &\sqrt{1+A^2\times\left(1-\frac{4}{D_r^4}\times\sum_{i,j} n_{ij}\left(D_r\left(C_{ij}-D_{ij}\right)-\left(P-Q\right)\left(N-n_i\right)\right)^2\right)} \end{aligned} \tag{30}$$

and

$$\begin{aligned} ASE_0(D_2) &\le \sqrt{VAR(D)\times\left[1+A^2\times\left(1-VAR(D)\right)\right]} \\ &= \frac{2}{D_r}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij}-D_{ij}\right)^2-\frac{1}{N}\left(P-Q\right)^2}\times \\ &\sqrt{1+A^2\times\left(1-\frac{4}{D_r^2}\times\sum_{i,j} n_{ij}\left(C_{ij}-D_{ij}\right)^2-\frac{1}{N}\left(P-Q\right)^2\right)} \end{aligned} \tag{31}$$

The former is used in estimating confidence intervals and the latter when testing null hypotheses.

Numerical example

As a numeric example of calculating G_2 and related ASEs, and confidence interval, let us use a polytomous dataset with N = 25 cases as in Table 3 adapted from Cox (1974, p. 177) and used by Drasgow (1986, p. 70). Let us assume that the variables are related to item g and score X.

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

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

Used by permisson of Biometric society

Table 4. Contingency table based on Table 3

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

For calculating G and G_2, based on Eq. (1),

$$C_{ij} = \sum_{h<i}\sum_{k<j} n_{hk} + \sum_{h>i}\sum_{k>j} n_{hk} = 480 \text{ ,}$$

$$D_{ij} = \sum_{h<i}\sum_{k>j} n_{hk} + \sum_{h>i}\sum_{k<j} n_{hk} = 420 \text{ ,}$$

$$P = \sum_{i,j} n_{ij} C_{ij} = 180 \text{ ,}$$

$$Q = \sum_{i,j} n_{ij} D_{ij} = 114 \text{ ,}$$

and, hence,

$$G = \frac{P-Q}{P+Q} = \frac{180-114}{180+114} = 0.224 \cdot$$

For $df(g) = 2$,

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

and, then,

$$G_2 = G\times\left(1+\left(1-G\right)\times A\right) = 0.224\times\left(1+\left(1-0.224\right)\right)\times 0.125 = 0.246 \text{ .}$$

As benchmarks, the estimate of the observed association by PMC is *Rit* = 0.185 and of the inferred association by the polychoric correlation R_{PC} = 0.123, although the latter value depends on the estimation method to some extent, and Somers' $D(g|X)$ = 0.219, and corrected D_2 = 0.240. Notably, the corrected form of D_2 gives estimates that are closer the value zero than the original, uncorrected form of D_2 (cf. D_2 = 0.317 in Metsämuuronen, 2020b).

For the ASEs,

$$\sum_{i,j} n_{ij}\left(QC_{ij} - PD_{ij}\right)^2 = 29,021,688 \text{ ,}$$

$$\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}\left(P-Q\right)^2 = 1534 - 174.24 = 1359.76 \text{ ,}$$

$$\frac{4}{\left(P+Q\right)^2} = \frac{4}{\left(180-114\right)^2} = 4.6277\times 10^{-5} \text{ ,}$$

and

$$\frac{2}{\left(P+Q\right)} = \frac{2}{\left(180+114\right)} = 0.00680 \text{ .}$$

Then,

$$ASE_1(G) = \frac{4}{\left(P+Q\right)^2}\sqrt{\sum_{i,j} n_{ij}\left(QC_{ij} - PD_{ij}\right)^2} = 4.6277\times 10^{-5}\times\sqrt{29,021,688} = 0.2493 \text{ ,}$$

$$ASE_0(G) = \frac{2}{\left(P+Q\right)}\sqrt{\sum_{i,j} n_{ij}\left(C_{ij} - D_{ij}\right)^2 - \frac{1}{N}\left(P-Q\right)^2} = 0.00680\times\sqrt{1359.76} = 0.2508 \text{ ,}$$

and, consequently,

$$ASE_1(G_2) \le \sqrt{VAR(G)\times\left[1+A^2\times\left(1-VAR(G)\right)\right]}$$
$$= \sqrt{0.2493^2\times\left[1+0.125^2\times\left(1-0.2493^2\right)\right]} = 0.2511 \ ,$$

and

$$ASE_0(G_2) \le \sqrt{VAR(G)\times\left[1+A^2\times\left(1-VAR(G)\right)\right]}$$
$$= \sqrt{0.2508^2\times\left[1+0.125^2\times\left(1-0.2508^2\right)\right]} = 0.2527 \ .$$

Both G and G_2 estimate the theoretical probability γ. Then, the traditional asymptotic 95% confidence interval for the true γ by G is

$$\gamma = G \pm t_{\alpha 0.975}(24)\times ASE_1(G) = 0.224 \pm 2.391\times 0.2493 = \left[-0.372, 0.820\right]$$

and by G_2

$$\gamma = G_2 \pm t_{\alpha 0.975}(24)\times ASE_1(G_2) = 0.246 \pm 2.391\times 0.251 = \left[-0.354, 0.846\right] .$$

Because Eq. (23), the factual interval in the latter interval is narrower than given here. The asymptotic significance can be approximated by Z test statistic $Z = \dfrac{G_2 - \gamma}{ASE_0(G_2)}$. When testing the hypothesis $\gamma = 0$,

$$Z = \frac{G}{ASE_0(G)} = \frac{0.246}{0.2527} = 0.973$$

leading to conclude that, given Table 3, the true γ could be zero (p = 0.165). This is indicated also by the confidence interval; zero belongs to the interval.

General characteristics of G_2

G_2 behaves according to the six principles set for correction. First, the estimates by G_2 tend to be higher than those by PMC (see Figure 6). Second, G_2 does not correct G when item discrimination is deterministic and G = 1 or G = 0. Third, the estimates by G are not corrected when $df(g)$ = 1. Fourth, the higher $df(g)$ is the greater the correction is in G_2. Fifth, G_2 does not produce out-of-range values. Of the 11,160 items on the simulation, none showed a value that was out of range regarding the limits of correlation. Notably, the magnitudes of the estimates by G_2 and R_{PC} are very close each other up to four categories in items after which the magnitudes of the estimates by D_2 are very close to R_{PC}.

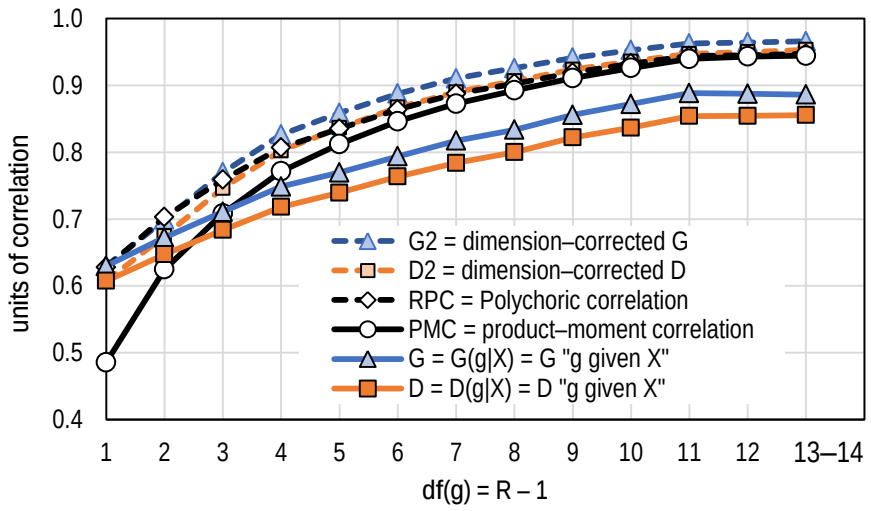


Figure 6. Average estimates of selected indices of IDP by varying df(g), k = 11,160 items

Overall, when it comes to correcting the obvious underestimation of association between an item and score in G, G_2 seems to behave logically at all levels of $df(g)$ used in the simulation. On average, G_2 underestimates the IDP remarkably

less than PMC, and notably less than *G*, as was the motivation for its derivation. The magnitude of the estimates by G_2 follows closely to that by the corrected D_2, which is expected because of the identical correction factor. Also, the variability in the magnitudes of the estimates by G_2 is smaller than that by *G* at each level of $df(g) > 1$.

Obvious underestimation and potential overestimation in G_2

A simple criterion for the obvious underestimation in estimates by G_2 is whether the magnitude of the estimates is lower than those by PMC. As a benchmark of G_2, the original *G* produced 2,435 such estimates (21.8%) where PMC > *G*. Notably, G_2 produced just 23 such estimates (0.2%). As a benchmark, the corrected D_2 produced 4,155 (37%) such estimates. It seems that the probability of obtaining obvious underestimation in real-life datasets is very low while using G_2.

With the condition of deterministic item discrimination, the value $G_2 = G = 1$ is accurate in reflecting the proportion of logically ordered test-takers in the item after they are ordered by the score (see Study 1). Also, if R_{PC} does not overestimate IDP when $df(g) = 1$, it is unlikely that $G_2 = G$ would obviously overestimate IDP with binary items (see Study 2). The magnitude of the estimates by G_2 is higher in comparison to that of D_2; this cannot be taken as an obvious overestimation because it is caused just by a different logic of calculating the probability.

Possible overestimation of IDP by G_2 in the polytomous case is not easy to evaluate in strict terms. If the magnitude of the estimates is higher than 1, those would be obvious overestimates. In the training and cross-validating datasets, none of the items exceeded 1. One potential criterion for overestimation is the "Guttman boundary" used by Metsämuuronen (2020b) when assessing a possible overestimation in D_2. Guttman boundary refers to the theoretical, maximally discriminating Guttman-patterned datasets, so-called Guttman scale (Guttman, 1950). In a Guttman-patterned dataset, *G* gives the maximal estimate 1 while the estimates by PMC are always smaller than 1. Assuming a score without ties, in the binary case, the highest value of *Rit* approximates $\rho_{gX}^{\max} = 0.866$ (see Metsämuuronen, 2020b; see also the latter set of Figures in Appendix) and, hence, the lowest point of the difference is $G - \text{PMC} = 1 - 0.866 = 0.134$ (see Figure 7).

The values of *G* – PMC or G_2 – PMC that exceed the Guttman boundary strictly indicate that the magnitudes of the estimates by *G* or G_2 and PMC are unexpectedly far from each other. However, this does not necessarily mean that the estimates by G_2 are overestimated; it may also indicate that the estimates by PMC are radically underestimated or that, in these cases, probability as an indicator of IDP detects more effectively the discrimination power in comparison with the covariance by PMC. Anyhow, as a rough tool, the Guttman boundary may indicate some latent behaviour of G_2 in comparison with *G*.

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

$$\frac{(X - x_0)^2}{a^2} + \frac{(Y - y_0)^2}{b^2} = \frac{(p - 0.5)^2}{0.5^2} + \frac{(Rit - 0)^2}{0.866^2} = 1 \qquad (32)$$

(Metsämuuronen, 2020b), where 0.866 refers to the limit of the maximum value of PMC in the deterministic pattern in the binary dataset. From (32) we solve *Rit*:

$$\text{PMC} = Rit = \sqrt{\left(1 - \frac{(p - 0.5)^2}{0.5^2}\right) \times 0.866^2} \qquad (33)$$

and, then, with the deterministically discriminating items,

$$G - \text{PMC} = 1 - \sqrt{\left(1 - \frac{(p - 0.5)^2}{0.5^2}\right) \times 0.866^2}\,. \qquad (34)$$

Guttman boundary is illustrated in Figure 7 where, notably, the asymmetry in the distribution points to a limitation in the original dataset: in the dataset with $\overline{p} > 0.60$, it is more probable to obtain extremely easy items with high *p* than extremely difficult items with low *p*.

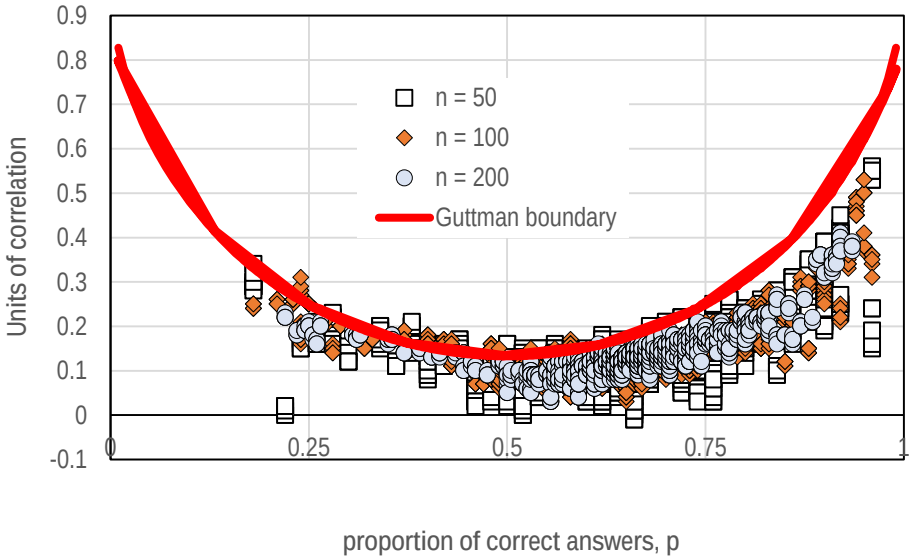


Figure 7. Guttman-pattern as a limit for the possible overestimation in G_2; G_2 *– PMC by* p*;* df(g) *= 1,* k *= 5 943 estimates*

In the training datasets, 131 out of 11,160 estimates by *G* (1.2%) exceeded the Guttman boundary, all with *df*(*g*) = 1 (Figure 8). G_2 produces one additional estimates of this kind totalling up to 132 cases (1.2%). The magnitude of the possible overestimation ranges 0–0.047 units of correlation with the average of 0.013.

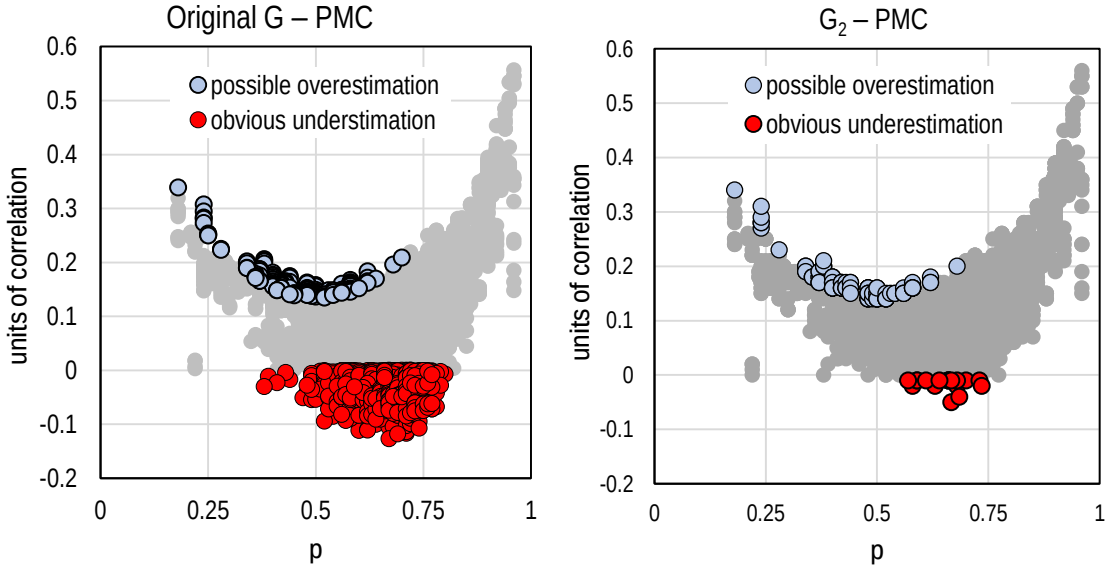


Figure 8. Possible overestimation and obvious underestimation in G and G_2

Discussion

The aim of the article was to study the characteristics of Goodman–Kruskal gamma in measurement modelling settings. *G* in general is very versatile when it comes to estimating the association between two variables with ordinal scale in educational settings (see, e.g., Good, 2015; Higham et al., 2016; Metsämuuronen, 2017; Shafina, 2021). By knowing the hidden directional nature of *G* to the same direction as we usually use eta squared in GLM settings (Metsämuuronen, 2021), it is good to be careful in its interpretation in the case that the scales of the variables differ from each other.

In measurement modelling settings related to educational realm, for example in item analysis of achievement or attitude tests, *G* appears to be a superior alternative for *Rit* and *Rir* with binary items and with polytomous items up to four categories. *G* appeared to be unexpectedly reliable as an estimator to produce accurately the latent perfect association between the item and the score irrespective of the distributions, degrees of freedom, the number of tied pairs and tied values in the variables, or the difficulty levels in the items. From this viewpoint, *G* carries the same characteristics as polychoric correlation coefficient although, from the computational viewpoint, R_{PC} is notably more complicated to calculate. However, *G* in comparison with *Rit*, tends to underestimate IDP in an obvious manner when the number of categories in the item exceeds four. Hence, a dimension–corrected *G*, G_2, was derived for *G*.

Based on the sub-studies in this article, some advantages of G and G_2 in relation to *Rit*, *Rir*, R_{PC}, D and D_2 in measurement modelling settings are the following (cf. Metsämuuronen, 2020a, 2020b):

1. G and G_2 are accurate in reflecting the latent perfect association between the item and the score unlike D, *Rit*, and *Rir*, because the last is based on the same mechanics as *Rit*. In this respect, G and G_2 have the same character as R_{PC}.
2. G and G_2 reach the values ±1 accurately, while *Rit* and *Rir* cannot reach the limits of correlation and R_{PC} cannot reach the extreme value with standard procedures.
3. G and G_2 are more robust for extreme observations, nonlinearity, and difficulty levels of the item than *Rit* and *Rir* because of being based on ranks.
4. G and G_2 are superior to *Rit* and *Rir* with dichotomous (G and G_2) as well as polytomous items (G_2) because, most probably, they produce an estimate that underestimates IDP *less* than *Rit* and *Rir* and, to some extent, than R_{PC}.
5. G and G_2 utilize the known composite of items and score while R_{PC} refers to unknown, unreachable, and hypothetical variables that are difficult to use in further research.
6. G and G_2 are applicable and accurate with non-normal datasets as well as sparse, small, or large cross-tables, while the applicability and accuracy of the estimation result of the *Rit* and *Rir* depend on the normality of the phenomenon.
7. G and G_2 have a logical directional nature from the measurement modelling viewpoint; they indicate how well the latent trait (score) explains the responses in the manifested variable (item).
8. G and G_2 make it possible to detect the maximally discriminating test items while *Rit* and *Rir* cannot detect this condition with real-life settings.
9. G and G_2 are reasonably easy to calculate, even manually, in practical test settings, while the calculation of R_{PC}, for example, requires complex procedures and specific software packages.
10. G and G_2 reach the meaningful direction of association strictly while D gives three options.
11. G and G_2 underestimate IDP less than D and D_2 do.

Although D and D_2 are offered as appealing alternatives to *Rit* and *Rir* in item analysis settings by Metsämuuronen (2020a; 2020b), it seems that G and G_2 may be even better options; G and G_2 underestimate IDP even less than D and $D2$. At least, the estimates by G and G_2 are less conservative in comparison with those by D and D_2.

Further possibilities of G and G_2 in educational measurement

PMC is an important referential coefficient of G and D, not only because it gives the benchmark for obvious underestimation for the estimates of IDP but because the underestimation in the *reliability* of the test score has been connected to the mechanical underestimation of true correlation by PMC. Metsämuuronen (2016), specifically, discussed the possibility to replace the range-restricted PMC with a "superior alternative" to PMC in the formulae of reliability to reduce the mechanical error in the estimates of reliability. G, and G_2 could be these options.

As an example, consider the most used estimator of reliability, coefficient alpha (Cronbach, 1951; Kuder & Richardson, 1937; see the frequent use of coefficient alpha, as examples, in Cheng et al., 2012; Green & Young, 2009; Trizano-Hermosilla & Alvarado, 2016). Coefficient alpha can be expressed in a form

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

(Lord & Novick, 1968) where k is the number of items, σ_g^2 is the item variance, and ρ_{gX} is *Rit* = PMC. While knowing that *Rit* underestimates the association between an item and score because of technical reasons, and G and G_2 underestimate this association less, we could use a "dimension-corrected" alpha or "systematic mechanical error corrected" or, "SME-corrected" alpha such as

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

with binary items and

$$\alpha_{G_2} = \frac{k}{k-1}\left(1 - \frac{\sum_{g=1}^{k}\sigma_g^2}{\left(\sum_{g=1}^{k}\sigma_g G_{2gX}\right)^2}\right) \quad (37)$$

with polytomous and mixed items where the quantity of SME may be remarkably reduced. Metsämuuronen and Ukkola (2019) used this kind of estimator when assessing reliability of the test scores related to the achievement levels of academic school readiness in mathematics and mother language at the beginning of the first grade (n > 7,000)—instead of G and G_2 they used D and D_2 in Eqs. (36) and (37). The difference between the traditional alpha (α_R) and the "SME-corrected" alpha with D and D_2 (α_D) was the clearest in tests where the item variances were extreme, that is, with extremely easy and extremely difficult tests. This is caused by the fact that when the test difficulty is extreme, either very easy or difficult, PMC is severely affected by SME, while D and G are much more robust and stable. In the dataset, the lowest estimates of reliability by the traditional alpha were α_R = 0.25, 0.43, 0.46, and 0.49, which were, by using a SME-corrected alpha, α_D = 0.86, 0.62, 0.66, and 0.81, respectively. This indicates that although the traditional coefficient alpha indicates that the scores were not able to discriminate between the test-takers, the SME-corrected estimates indicate opposite: systematically, the higher achieving pupils were able to produce correct answers more probably than the lower-achieving pupils. Hence, factually, the score was able to discriminate between the test-takers decently if not highly. This means that even though the traditional alpha may doom a test to be undiscriminating, this may be caused by a technical error in estimating item–total correlation; the "real" reliability could be notably higher. If used G and G_2 instead of D and D_2, the magnitudes of the estimates in Metsämuuronen and Ukkola (2019) would have been somewhat higher because the values by G are higher than those by D. The characteristics of these kinds of estimators are not discussed here and further studies in this area would be worth conducting.

Another option to study further relates to a somewhat surprising by-result of the simulation with perfectly correlating latent variables. The simulation showed that G and R_{PC} were the only ones in comparison that share the character of producing (correctly) the perfect correlation irrespective of the distributions of the item and the score, number of cases, degrees of freedom of the item and the score, the number of tied pairs and tied values in the variables, or the difficulty levels in the items. Then, it seems that G and R_{PC} may have some common character yet unrevealed.

Conclusions

All in all, although Goodman–Kruskal gamma (G) is used relatively rarely in educational settings in comparison with PMC, it has promising potential as a coefficient of association. G can be used whenever two variables are measured, at least, in an ordinal or interval scale, and it estimates the probability that cases are in same order in two variables and, on the other hand, it indicates the proportion of logically ordered cases in a variable with a narrower scale after they are ordered by the variable with wider scale. In the item analysis settings, G appears to be unexpectedly appealing as an estimator of association between an item and a score because it indicates the probability to get a correct answer in the test item given the score, and it accurately produces perfect latent association irrespective of distributions, degrees of freedom, number of tied pairs and tied values in the variables, or the difficulty levels in the items.

Because of carrying the same deficiency as Somers' D and Kendall's *tau* to underestimate the item discrimination power in an obvious manner when the number of categories increases in the item, a simple transformation of G, "dimension-corrected G" (G_2) is proposed to be used in the measurement modelling settings. Both G and G_2 appear to be promising alternatives to item–total correlation and item–rest correlation coefficient in measurement modelling settings, G with binary items and G_2 with binary, polytomous and mixed datasets. They can be used strictly in item analysis to indicate the item discriminating power and they could be used in estimating reliability of the score. Specifically, in the case that the test items are very easy or very difficult to the test-takers, new estimators named "SME-free estimators of reliability" that use G or $G2$ (or D and D_2) instead of PMC in the estimators were introduced. These estimators may reveal that a measurement instrument doomed to be poor by the traditional coefficient alpha because of PMC embedded in the coefficient may, in fact, discriminate between the test-takers remarkably better than expected.

Limitations

One obvious challenge in generalizing the new coefficient is that G_2 is *developed for item analysis settings*. In these settings, always $R << C$, and the items and the score are mechanically connected. Notably, the dimension correction leads, automatically, to approximate the perfect value $G_2 = 1$ (or, in the ultimate pathological case, to $G_2 = -1$) when the item is a continuous one and the sample size is large. Because of this, the applicability of G_2 may be reduced to measurement modelling settings with items that have a narrow scale and it will not be wise to use G_2 as a general coefficient without further studies and possible amendments. From this perspective it would be beneficial to compare G_2 with the other corrections suggested for G (e.g., Bai & Wei, 2009; Highan & Higham, 2019; Hryniewicz, 2006; Kvålseth, 2017; Masson & Rotello, 2009; Rousson, 2007).

Second, the *benchmark of the possible underestimation in G was PMC* while, perhaps, a coefficient called *r*-polyreg correlation, an *r*-polyserial estimated by regression correlation (Livinstone & Dorans, 2004) would be a more proper benchmark. This coefficient, developed to overcome the challenge of the obvious overestimation in biserial and polyserial correlation coefficient, does not exceed 1, nor does it rely on bivariate normality assumptions (see Moses, 2017). More studies may be valuable in this respect.

Third, G_2 is based on empirical items that embed the *limitations of the original datasets* to a certain extent. We do not know how much the estimates depend on the original dataset. From the cross-validating dataset it is known that when the scale of the score gets higher the model created by the training dataset gives conservative estimates with longer tests than $df(X) > 28$. There are no numerical sub-coefficients in the correction factors in Eqs. (10) and (11) although related to the original dataset. Hence, to some extent, G_2 is more general than when it includes specific numerical coefficient(s) that are strictly dependent on the underlying dataset. Simulation with $df(g) > 6–7$ would enrich our knowledge of the applicability of G_2.

Acknowledgements

The writer sincerely thanks Dr. Roger Newson, research associate at the Faculty of Medicine, School of Public Health, Imperial College, London, for leading 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 IDP in both *D* and *G*. Sincere thanks also to Counsellor of Evaluation Jukka Marjanen at FINEEC who solved a practical challenge in calculating ASEs of *G* and G_2.

References

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

Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London, 27*(325–336), 186–190. http://doi.org/10.1098/rstl.1710.0011

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

Bai, J., & Wei, L.-L. (2009). A new method of attribute reduction based on gamma coefficient. In S.-M. Zhou & W. Wang, *GCIS 2009. 2009 WRI Global Congress on Intelligent Systems* (pp. 370–373). IEEE Computer Society. https://doi.org/10.1109/GCIS.2009.212

Bravais, A. (1844). *Analyse Mathematique. Sur les probabilités des erreurs de situation d'un point* [Mathematical analysis. On the probabilities of the point errors]. Imprimerie Royale.

Breslow, N. (1970). A generalized Kruskal–Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrics/ Biometrika, 57*(3), 579–594. http://doi.org/10.1093/biomet/57.3.579

Byrne, B. M. (2016). *Structural Equation Modeling with AMOS. Basic concepts, applications, and programming* (3rd ed.). Routledge.

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

Cleff, T. (2019). *Applied Statistics and Multivariate Data Analysis for Business and Economics. A Modern Approach Using SPSS, Stata, and Excel*. Springer.

Conover, W. J. (1980). *Practical nonparametric statistics*. Wiley & Sons.

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

Davis, J. A. (1967). A partial coefficient for Goodman and Kruskal's gamma. *Journal of the American Statistical Association, 62*(317), 189–193. https://doi.org/10.1080/01621459.1967.10482900

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

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

El-Shaarawi, A. H., & Piegorsch, W. W. (2001). *Encyclopedia of Environmetrics (Volume 1).* John Wiley and Sons.

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

Forthmann, B., Förster, N., Schütze, B., Hebbecker, K., Flessner, J., Peters, M. T., & Souvignier, E. (2020). How much *g* is in the distractor? Re-thinking item-analysis of multiple-choice items. *Journal of Intelligence*, *8*(1), 1-36. https://doi.org/10.3390/jintelligence8010011

Freeman, L. C. (1986). Order-based statistics and monotonicity: A family of ordinal measures of association. *Journal of Mathematical Sociology, 12*(1), 49–69. https://doi.org/10.1080/0022250X.1986.9990004

Galton, F. (1889). Kinship and correlation. *Statistical Science*, *4*(2), 81–86. http://doi.org/10.1214/ss/1177012581

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrics/ Biometrika, 52*(1–2), 203–233. http://doi.org/10.1093/biomet/52.1-2.203

Gini, C. (1912). *Variabilità e mutabilità. Contributo allo studio delle distribuzioni e dellerelazioni statistiche* [Variability and mutability. Contribution to the study of distributions and statistical relationships]. Bologna.

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

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores*. Psychological Bulletin, 119*(1), 159–165. https://doi.org/10.1037/0033-2909.119.1.159

Good, K. (2015). Investigating relationships between educational technology use and other instructional elements using "big data" in higher education [Doctoral dissertation, Iowa State University]. Iowa State University Digital Repository. https://lib.dr.iastate.edu/etd/14854

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

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classification*. Springer-Verlag.

Green S. B., & Yang Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrics/ Psychometrika, 74*(1), 121–135. http://dx.doi.org/10.1007/s11336-008-9098-4

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.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfield, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60 – 90)*.* Princeton University Press.

Harrell, F. (2001). *Regression Modelling Strategies*. Springer.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, *247*(18), 2543–2546. http://doi.org/10.1001/jama.1982.03320430047030

Heagerty, P. J., & 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

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

Higham, P. A., & Higham, D. P. (2019). New improved gamma: Enhancing the accuracy of Goodman-Kruskal's gamma using ROC curves. *Behavior Research Methods*, *51*(1), 108–125. https://doi.org/10.3758/s13428-018-1125-5

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2016). Internal mapping and its impact on measures of absolute and relative metacognitive accuracy. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.15

Hryniewicz, O. (2006). Goodman-Kruskal γ measure of dependence for fuzzy ordered categorical data. *Computational Statistics & Data Analysis, 51*(1), 323–334. https://doi.org/10.1016/j.csda.2006.04.014

IBM (2017). *IBM SPSS Statistics 25 Algorithms*. IBM.

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

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

Kendall, M. G. (1948). *Rank correlation methods* (1st ed.). Charles Griffin & Co. Ltd.

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

Kim, J.-O. (1971). Predictive measures of ordinal association. *American Journal of Sociology, 76*(5), 891–907. https://doi.org/10.1086/225004

Kreiner, S., & Christensen, K. B. (2009). Item screening in graphical loglinear Rash models. *Psychometrics/ Psychometrika, 76*(2), 228–256. https://doi.org/10.1007/s11336-011-9203-y

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks on one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621. http://doi.org/10.2307/2280779

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

Kvålseth, T. O. (2017). An alternative measure of ordinal association as a value-validity correction of the Goodman–Kruskal gamma. *Communications in Statistics - Theory and Methods, 46*(21), 10582–10593. https://doi.org/10.1080/03610926.2016.1239114

Livingston, S. A., & Dorans, N. J. (2004). *A graphical approach to item analysis* (Research Report No. RR-04-10). Educational Testing Service. http://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.

Love, T. E. (1997). Distractor selection ratios. *Psychometrics/ Psychometrika,* 62(1)**,** 51–62. https://doi.org/10.1007/BF02294780

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrics/ Econometrica, 13*(3), 245–259. https://doi.org/10.2307/1907187

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*(1), 50–60. http://doi.org/10.1214/aoms/1177730491

Martin, W. S. (1973). The effects of scaling on the correlation coefficient: A test of validity. *Journal of Marketing Research*, *10*(3), 316–318. http://doi.org/10.2307/3149702

Martin, W. S. (1978). Effects of scaling on the correlation coefficient: Additional considerations. *Journal of Marketing Research, 15*(2), 304–308. https://doi.org/10.1177/002224377801500219

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 509–527. https://doi.org/10.1037/a0014876

McDonald, R. P. (1985). *Factor analysis and related methods*. Lawrence Erlbaum Associates.

Meade, A. W. (2010). Restriction of range. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1278–1280). SAGE Publications, Inc. http://doi.org/10.4135/9781412961288.n309

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

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

Metsämuuronen, J. (2020a). Somers' *D* as an Alternative for the Item–Test and Item–Rest Correlation Coefficients in the Educational Measurement Settings. *International Journal of Educational Measurement, 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. *International Journal of Educational Methodology, 6*(2), 297–317. https://doi.org/10.12973/ijem.6.2.297

Metsämuuronen, J. (2021). 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. ResearchGate. http://doi.org/10.13140/RG.2.2.19404.44163

Metsämuuronen, J., & Ukkola, A. (2019). *Alkumittauksen menetelmällisiä ratkaisuja* [Methodological solutions of zero level assessment]. Finnish Education Evaluation Centre.

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. http://doi.org/10.1007/978-3-319-58689-2_2

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal, 2*(1), 45–64. https://doi.org/10.1177/1536867X0200200103

Newson, R. (2006). Confidence intervals for rank statistics: Somers' *D* and extensions. *The Stata Journal, 6*(3), 309–334. https://doi.org/10.1177/1536867X0600600302

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

Nielsen, J. B., Kyvsgaard, J. N., Sildorf, S. M., Kreiner, S., & Svensson, J. (2017). Item analysis using Rasch models confirms that the Danish versions of the DISABKIDS® chronic-generic and diabetes-specific modules are valid and reliable. *Health Qual Life Outcomes 15*(1)**,** article 44, 1–10. https://doi.org/10.1186/s12955-017-0618-8

Nielsen, T., & Santiago, P. H. R. (2020). Using graphical loglinear Rasch models to investigate the construct validity of Perceived Stress Scale. In M. S. Khine (Ed.), *Rasch Measurement: Applications in Quantitative Educational Research* (pp. 261–281). Springer Nature. https://doi.org/10.1007/978-981-15-1800-3_14

Olsson, U. (1980). Measuring correlation in ordered two-way contingency tables. *Journal of Marketing Research, 17*(3), 391–394. https://doi.org/10.1177/002224378001700315

Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution.- III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society A, 187,* 253–318. https://doi.org/10.1098/rsta.1896.0007

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

Raykov, T., & Marcoulides, G. A. (2013). Meta-analysis of reliability coefficients using latent variable modeling. *Structural Equation Modeling*, *20*(2), 338–353. http://doi.org/10.1080/10705511.2013.769396

Rousson, V. (2007). The gamma coefficient revisited. *Statistics & Probability Letters 77*(17), 1696–1704. https://doi.org/10.1016/j.spl.2007.04.009

Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effect of range restriction on predictor intercorrelations. *Journal of Applied Psychology, 92*(2), 538–544. http://doi.org/10.1037/0021-9010.92.2.538

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Sen, P. K. (1963). On the estimation of relative potency in dilution(-direct) assays by distribution-free methods. *Biometrics, 19*(4), 532–552. https://doi.org/10.2307%2F2527532

Shafina, A (2021). The impact of birth-order, sib-size, siblings' sex composition on educational attainment in the Maldives. *The Universal Academic Research Journal, 3*(2), 87–100.

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.). Chapman & Hall/CRC.

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

Sirkin, M. R. (2006). *Statistics of the social science* (3rd ed.). SAGE Publications, Inc.

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

Somers, R. H. (1980). Simple approximations to null sampling variances. Goodman and Kruskal's gamma, Kendall's tau and Somers d*yx*. *Sociological Methods & Research, 9*(1), 115–126. https://doi.org/10.1177/004912418000900107

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Investigations of the mathematics/ Indagationes Mathematicae, 14*(3), 327–333. http://doi.org/10.1016/S1385-7258(52)50043-X

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis I, II, III. In *Proceedings of the Section of Sciences - Koninklijke Nederlandsche Akademie van Wetenschappen* [Royal Netherlands Academy of Sciences] (Series A. Mathematical Sciences, pp. 386–392, 521–525, 1397–1412). North-Holland.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7*, 1-8. https://doi.org/10.3389/fpsyg.2016.00769

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

Wholey, J., S., Hatry, H., P., & Newcomer, K. E. (Eds.) (2015). *Handbook of practical program evaluation* (4th ed.). Jossey-Bass.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. http://doi.org/10.2307/3001968

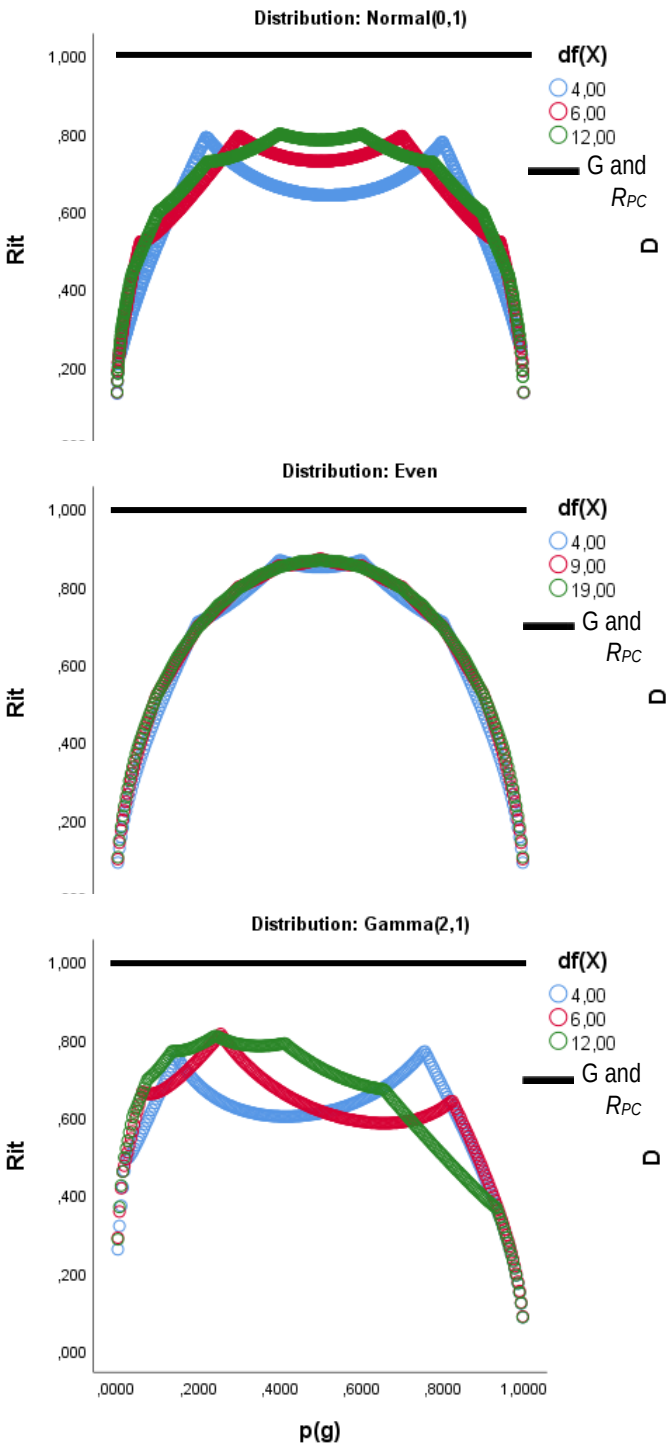
Wilson, T. P. (1974). Measures of association for bivariate ordinal hypotheses. In H. M. Blalock (Ed.), *Measurement in the social sciences* (pp. 327–342). Aldine.

Zaionts, C. (2020). *Polychoric correlation using solver*. Real Statistics Using Excel. http://www.real-statistics.com/correlation/polychoric-correlation/polychoric-correlation-using-solver/

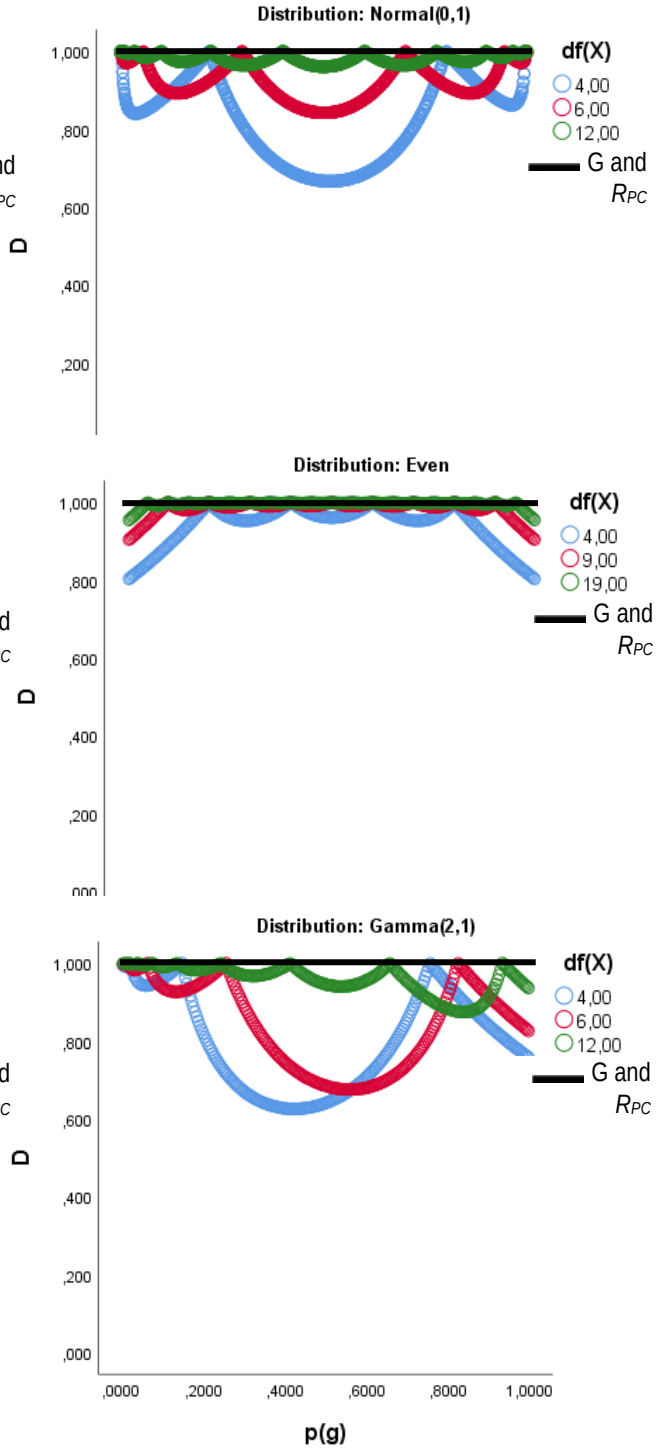
Appendix 1.

Estimates of IDP in Study 1

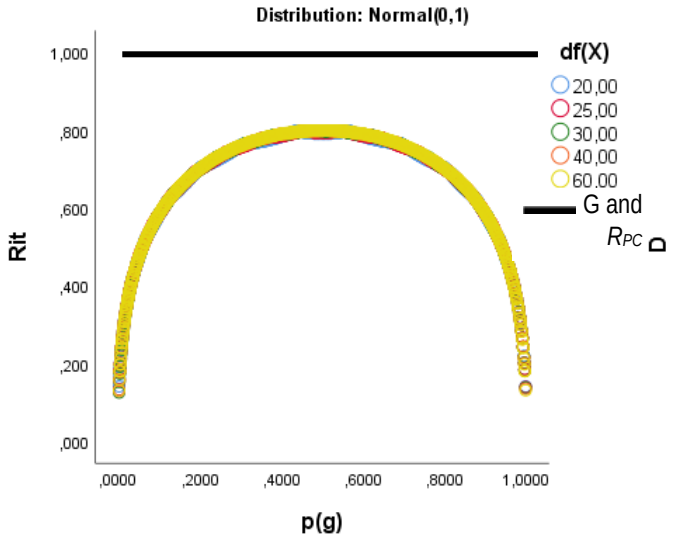
estimates by PMC, G, and R_{PC}

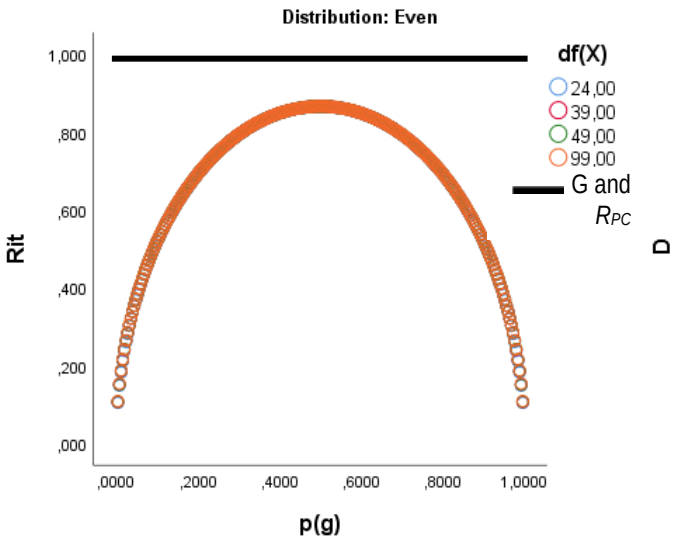


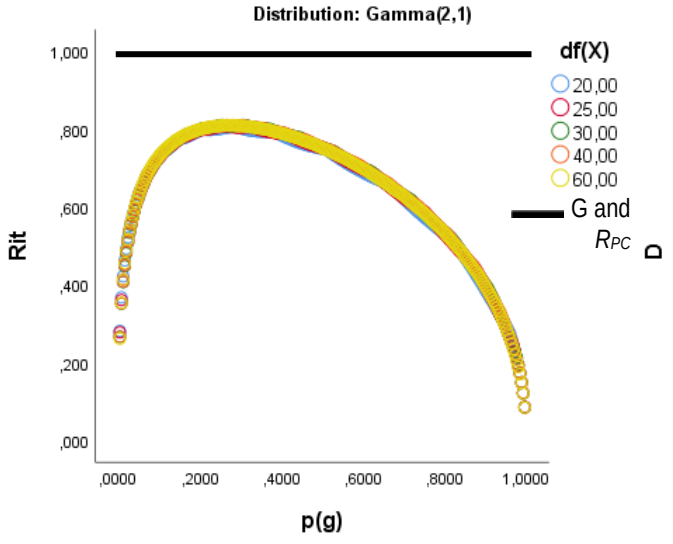
estimates by D, G, and R_{PC}



estimates by PMC, G, and R_{PC}







estimates by D, G, and R_{PC}