The Journal of Experimental Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/vjxe20

Five New Nonparametric Estimators of Common Language Effect Size

Jari Metsämuuronen

To cite this article: Jari Metsämuuronen (16 Feb 2025): Five New Nonparametric Estimators of Common Language Effect Size, The Journal of Experimental Education, DOI: 10.1080/00220973.2025.2459411

To link to this article: https://doi.org/10.1080/00220973.2025.2459411

THE JOURNAL OF EXPERIMENTAL EDUCATION
https://doi.org/10.1080/00220973.2025.2459411

Routledge
Taylor & Francis Group

MEASUREMENT, STATISTICS, AND RESEARCH DESIGN

Five New Nonparametric Estimators of Common Language Effect Size

Jari Metsämuuronen

University of Turku, Turku, Finland

ABSTRACT

Five new nonparametric effect size estimators are proposed: the "probability of higher subgroup dominance based on Somers' *D*" (*PHD*), and its variant *PHG*, based on Goodman-Kruskal gamma (*G*), and three new variants of traditional estimators, Grissom-Kim *PS*, Cliff's *d*, and Vargha-Delaney *A*, called *ordinal PS*, *ordinal d*, and *ordinal A*. *PHD* and *PHG* are general estimators in three respects. First, they can be used strictly in polytomous ordinal settings, since the embedded estimators *D* and *G* are not restricted to binary settings. Second, it is shown that in the dichotomous cases, *PS*, *d*, and *A*, are special cases of *D* and hence of *PHD*. Third, because *PHD* and *PHG* are based on established correlation estimators, their calculation is straightforward, and they can be readily used in meta-analyses, provided that the embedded correlation estimators, *D* and *G*, are typically included in the report. The asymptotic standard errors for *PHD* and *PHG* and *ordinal PS, d,* and *A* are given, and their advantages are discussed in relation to those of Cohen's *d*, *f*, and η^2. *PHD* can be used as the basis for common language interpretation of the effect size *r*, Cohen *d*, Cohen *f*, and certain reliability estimators.

KEYWORDS

Common language effect size; dominance statistic; effect size; Goodman-Kruskal gamma; probability of superiority; rank-biserial correlation; rank polyserial correlation; Somers delta; stochastic superiority

Introduction

Effect size (ES) is a concept related to measuring the strength of the relationship between two variables in a population, or a sample-based estimate of this quantity. Kelley and Preacher (2012) define ES as a quantitative reflection of the magnitude of a phenomenon used to address a question of interest. While the traditional inferential statistic (*p*) is strictly dependent on the sample size—a large sample size leads to small standard errors and low *p* regardless of the actual size of the association or difference in means—the magnitude of the estimates of different estimators of ES is (largely) independent of the sample size. However, as Metsämuuronen (2023a, 2023b; see also McGrath & Meyer, 2006; Ruscio, 2008) has shown, differences in the proportion of cases in different subpopulations can affect the results. Consequently, if the sample size is large enough, the association may be statistically significant, but the magnitude may be trivial. This was the main rationale behind the decision by leading journals in the field of empirical education and psychology, such as *Measurement and Evaluation in Counseling and Development*, *The Journal of Experimental Education*, and *Educational and Psychological Measurement*, to mandate that the inclusion of some form of measure of ES should be reported along with the statistical significance

CONTACT Jari Metsämuuronen jari.metsamuuronen@gmail.com Turku Research Institute for Learning Analytics, Faculty of Science, University of Turku, Turku 20014, Finland.

(see editorials by Thompson, 1988, 1989, 1993, 1994; see history and literature in Huberty, 2002; Peng & Chen, 2014). This has been encouraged by the American Psychological Association (APA) and the American Educational Research Association (AERA), which have recommended the reporting of ES (Wilkinson and Task Force on Statistical Inference, American Psychological Association, Science Directorate, 1999 onwards; AERA, 2006 onwards).

The concept of statistical significance was established in the 1920s and 1930s (see Fisher, 1925; Neyman & Pearson, 1933a, 1933b). In contrast, the concept of ES has led to several approaches or "families" of estimators with different characteristics. Huberty (2002) and Peng and Chen (2014) have typologized these. The family related to relationship-based ES includes estimators such as product-moment correlation (r), coefficient of determination (r^2, R^2), and eta squared (η^2) as well as related population estimators and corrected estimators (e.g., ω^2, ε^2). The family related to group differences includes Cohen's d, Cohen's f, and f^2, Glass' delta, and Hedge's g. Cohen's d and f can be placed in either family because they can be expressed in terms of either correlation or test statistic related to group differences. The family related to the group overlap includes Cohen's U and the Huberty-Lowman improvement-over-chance classification (I). The family related to categorical variables includes *phi* coefficient, Cramer's V, Cohen's w, Cohen's h, and Freeman's *theta*. The last can be used with nominal-ordinal settings.

Cohen's estimators are discussed in detail in Cohen's classic volume related to effect sizes (Cohen, 1969, 1988), and numerous other estimators are presented in later publications (e.g., Grissom & Kim, 2012). Of these, the two most frequently reported measures for ES appear to be r and d (Funder and Ozer, 2019; Schäfer & Schwarz, 2019; see also Peng & Chen, 2014).

Another family, particularly interesting in terms of the focus of the article and the increase in volume (see Figure 1), are the nonparametric estimators of ES. These can be based on the median absolute deviation as an alternative to Cohen's d (Ricca & Blaine, 2022), or on nonparametric correlation coefficients such as Somers' delta (D; Somers 1962) or Goodman-Kruskal gamma (G; Goodman & Kruskal, 1954) as an alternative to r effect size. The latter are the focus in this article. They lead to the so-called common language effect sizes (McGraw & Wong, 1992). This family of estimators includes the original "common language" estimator (*CL*; McGraw & Wong, 1992), the "probability of superiority" (*PS*; Grissom, 1994; Grissom & Kim, 2001), "dominance statistic" also known as Cliff's d (Cliff, 1993), and "stochastic superiority" also known as Vargha-Delaney A (Delaney & Vargha, 2002; Vargha & Delaney, 2000). However, the epithets do not seem to be fixed; Ruscio and colleagues (e.g., 2008, 2012, 2013, 2022; see also Ortelli, 2022) use the term "probability of superiority" when referring to Vargha-Delaney A. It is shown that, in the

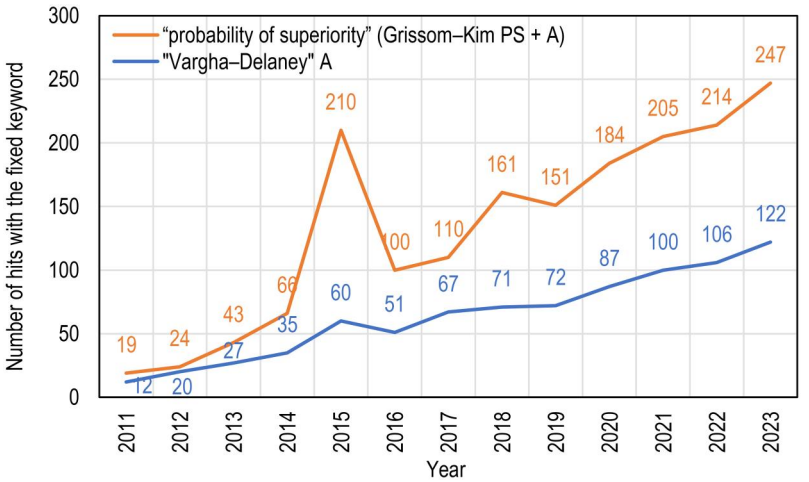

Figure 1. Number of hits of fixed keywords in Google Scholar related to selected nonparametric effect size estimators.

dichotomous settings, the formulas of *PS* and *A* are algebraically identical, although they use slightly different notation (see Appendix 1). Therefore, referring to both as "probability of superiority" is not a misnomer. In this article, the epithets are used in the same way as in Peng and Chen (2014).

Of the traditional nonparametric estimators, *A* has the widest applicability when it comes to multiple comparisons in the polytomous setting. Vargha and Delaney (1998, 2000) discuss advances in the traditional uses of *CL*, i.e., the two-group case, the *n*-group case, the two-group correlated sample case, and the discrete population case. Ruscio and Gera (2013) extended *A* with a variant to compare the subpopulations in ordinal settings (A_{ord}), a broader use of *A* to compare a single group with the union of the other subpopulations (A_{ik}), and to compare weighted datasets. Ruscio and Mullen (2012) provided confidence intervals for *A*.

Unlike parametric estimators such as *r*, R^2, η^2, Cohen's *d* and *f*, the nonparametric alternatives do not require assumptions about the population distribution. Grissom and Kim (2001) further noted that these nonparametric alternatives reconceptualize ES as an effect throughout a distribution, not just at its center. It may also be easier to communicate the meaning of the results with nonparametric estimates, hence the name 'common language' effect size (McGraw & Wong, 1992).

Peng and Chen (2014) observed that nonparametric estimators are rarely and briefly discussed even in specific textbooks on ES (see a literature review in Peng & Chen, 2014; however, see Grissom & Kim, 2012). Compared to the dominant estimators, such as η^2 and Cohen's *d*, even the most frequently cited nonparametric options, Grissom-Kim *PS* and Vargha-Delaney *A*, seem to be used relatively infrequency in practice. However, according to a nonsystematic and rough study conducted on Google Scholar at the time the article was finalizing (Nov. 27, 2024), a linear trend in the popularity of the estimators in publications is observed (Figure 1). One possible explanation for this phenomenon is the availability of relevant R libraries that include Vargha-Delaney's *A* and Cliff's *d*. For example, Torchiano's (2020) package *effsize* includes both *A* and *d*, while Ruscio's (2022) package *RProbSub* includes different versions of *A*, but under the name "probability of superiority," which corresponds to *PS*. Given their growing prevalence, it is prudent to conduct a more comprehensive study of these estimators.

Research problems and the course of study

Grissom-Kim *PS,* Cliff's *d*, and Vargha-Delaney *A* appear to have two major shortcomings. First, *A*, *d*, and *PS* were developed for and used primarily in binary or dichotomous settings as the nonparametric options to Cohen's *d* (see Cliff, 1993; Grissom & Kim, 2001; Vargha & Delaney, 2000). Although Vargha and Delaney (1998, 2000) have extended *A* to multiple groups and correlated data (see summary in Ruscio & Gera, 2013), the following numerical example illustrates that the interpretation of *A* in polytomous ordinal cases does not lead to a "common language" interpretation, as discussed by McGraw and Wong (1992). Since the basic form of the estimators (see Peng & Chen, 2014) can be expressed by using the Wilcoxon-Mann-Whitney *U* statistic (see Appendix 1), they are, basically, limited to two categories. This may be sufficient for the simplest two-group experimental designs. Appendix 1 shows that, in the dichotomous settings, all these estimators are special cases or modifications of Somers *D* which is *not* restricted to only two categories.

Second, Peng and Chen (2014) note that a specific shortcoming of *PS*, *A*, and Cliff's *d* is that they are difficult to use in meta-analyses without providing the raw data if the estimates of *PS*, *A*, or Cliff's *d* are not reported. The reason is that if the estimates are not published, unlike with *r*, η, Cohen's *d*, or Cohen's *f* which can be converted to each other's scale, "it is impossible to synthesize estimates […], if primary studies do not provide raw data from which to compute these

estimates" (p. 41). In this regard, it should be noted that the calculation of Cohen's d is feasible in cases where the values of r, η, or f are made available.

Primarily, two new nonparametric estimators of ES are proposed: estimators of the "probability of higher subgroup dominance" based on Somers' D (*PHD*) and Goodman-Kruskal gamma (G; *PHG*), and their advantages and shortcomings are discussed in relation to their dominant counterparts, r, Cohen's d and f as well as R^2, and η^2. Secondarily, it is shown that the conventional *PS*, d, and A can be expressed using either *PHD* or D in the dichotomous settings. This leads to a novel solution for *PS*, d, and A in the case of polytomous ordinal settings. In light of this, three generalizations and modifications of the traditional estimators are proposed: "ordinal *PS*," "ordinal d," and "ordinal A." Asymptotic standard errors are derived for all five estimators.

The study begins with the justification of five novel effect size estimators in Section ""Probability of the higher subgroup dominance", as an estimator of common language effect size". A numerical example is presented in Section "Numerical examples of the ordinal CLESs". Their asymptotic standard errors are derived in Section "Asymptotic standard errors of the ordinal CLESs". Their relation to Cohen's d and related parametric estimators of ES is discussed in Section "Empirical relation of PHD & PHG with Cohen's d".

"Probability of the higher subgroup dominance", as an estimator of common language effect size

In the conventional approach to *PS*, d, and A, the subpopulations within the categorical variable are represented by the values 1 and 2, respectively. In the following sections, the convention used in experimental studies, namely dummy variables and measurement modeling with test items, will be used. In this approach, the groups or subpopulations are denoted by 0 and 1 or 0, 1, 2, and so on. Group 1 is considered to be the "higher" group or subpopulation with respect to group 0, but "lower" with respect to group 2. In binary settings, this logic allows us to understand why the subpopulation denoted by 1 should be considered the "dominant" or "favored" group over the subpopulation denoted by 0. In most cases, the group of interest (such as the experimental group, the dummy subpopulation, or the cases that provided the correct answer to a test item) is labeled 1, while the remaining groups (the control group, the "other" cases, or the cases that gave an incorrect answer) are labeled 0. However, shifting the scale has no effect on the treatment.

The methodology and the basis for the nonparametric effect sizes discussed below are based on Cureton's (1956) concept of effect size and its relationship with the Mann–Whitney–Wilcoxon U statistic, Jonckheere–Terpstra JT statistic, and Somers D. These are examined in Sections "Cureton's effect size" & "Relation of Cureton's effect size & Somers' delta". A novel estimator of ES, based on Somers' delta, is presented in Section "PHD, "Probability of higher subgroup dominance" based on Somers D". Section "Three novel CLES estimators for ordinal settings based on Somers' D" discusses the generalizations and modifications of *PS*, d, and A, based on their relation to Somers D in the dichotomous settings. Section "PHG, "Probability of higher subgroup dominance" based on Goodman–Kruskal G" discusses a corresponding counterpart based on Goodman-Kruskal gamma. Section "Numerical examples of the ordinal CLESs" gives numerical examples.

Cureton's effect size

The effect size associated with Cureton's (1956) rank-biserial correlation (R_{RB}) is based on the idea of the proportion of favorable cases (f) and unfavorable (u) cases:

$$R_{RB} = f - u = f - (1 - f) = 2f - 1 \tag{1}$$

(see Berry et al., 2018). This is well related to the logic used in computing the Wilcoxon-Mann-Whitney U statistic (see, e.g., Siegel & Castellan, 1988). Recall that the U statistic takes two values

(U_0 and U_1), the *lower* of which is used in statistical inference. The basic logic of Cureton's effect size leads us to use the *higher-ranked* subpopulation statistic (U_1) as the basis for the effect size; U_1 refers to the number of pairs of observations in which the cases in subpopulation 1 rank higher than the cases in subpopulation 0 after being ordered by a metric variable with an ordinal, interval, semicontinuous, or continuous scale. Using this notation, the proportion of favorable cases is

$$f = \frac{U_1}{n_0 n_1} \tag{2}$$

where $n_0 n_1$ refers to the number of all pairs in the data set. Consequently, as seen in Equation (1), Cureton's R_{RB} is formed by doubling and relocating this entity:

$$R_{RB} = 2 \times \frac{U_1}{n_0 n_1} - 1 \tag{3}$$

The same result can be obtained using Wendt's formula (1972), which is based on the U-value related to the *lower* of the subpopulations (U_0):

$$R_{RB} = 1 - 2 \times \frac{U_0}{n_0 n_1} \tag{4}$$

Newson (2008; see also Metsämuuronen, 2021a) showed that Cureton's R_{RB} is a special case of Somers D for the dichotomous settings and directional so that the order in the grouping variable (g) depends on the order of the metric variable (X), $D = D(g|X)$, "g given X", which is usually labeled as "X dependent" in the software packages (see the discussion regarding a proper direction in Metsämuuronen, 2020 and of the unconventional notation of $g|X$ in Metsämuuronen, 2021a, 2021b). Metsämuuronen (2021a, 2022a) generalized the same for the polytomous ordinal settings. The relationship between Cureton's effect size and Somers' delta is discussed in below.

Relation of Cureton's effect size and Somers' Delta

Assume two random variables with dichotomous, binary, or ordinal polytomous scale in the categorical variable (later g) and a metric scale (later X) with observed values x_i and y_j, respectively. The observed values in g have categories $R = 1, \ldots, R$ and the observed values in X have categories $C = 1, \ldots, C$, and $R < C$. Let then (x_1, y_1), (x_2, y_2), $\ldots$, (x_N, y_N) be a set of observations from the variables g and X. The random pairs of observations (x_l, y_l) and (x_h, y_h), where l (from "lower") $< h$ (from "higher"), are concordant if the order for both elements matches, i.e., if $x_l < x_h$ and $y_l < y_h$ or $x_l > x_h$ and $y_l > y_h$. The number of concordant pairs in the given data set is denoted by P. Similarly, the pairs are discordant if $x_l < x_h$ and $y_l > y_h$ or $x_l > x_h$ and $y_l < y_h$ simultaneously. The number of discordant pairs is denoted by Q. In the direction relevant to the proposed estimators, i.e., $D = D(g|X)$, if $y_l = y_h$, the pairs are tied; they are neither concordant nor discordant.[1] Note that, in this direction, in all pairs, $x_l \neq x_h$, since the cases are selected from different categories.

The computational form of D directed so that "g given X" or "X dependent", i.e., $D(g|X) = D$, can be expressed as follows:

$$D = D(g|X) = \frac{2(P - Q)}{N^2 - \sum_{i=1}^{R}(n_i{}^2)} \tag{5}$$

(e.g., Siegel & Castellan, 1988), where N refers to the total number of cases and n_i refers to the number of cases in the subpopulations related to the categorical ordinal variable g, and $N^2 - \sum_{i=1}^{R}(n_i{}^2)$ refers to the number of all possible pairs in the given data set. Note that in Equation (5), only half of the Ps and Qs are computed, and they are doubled in the formula. N can be expressed as follows:

$$N = \sum_{i=1}^{R} n_i \tag{6}$$

Because of (6),

$$N^2 - \sum_{i=1}^{R} \left(n_i{}^2\right) = \left(\sum_{i=1}^{R} n_i\right)^2 - \sum_{i=1}^{R} \left(n_i{}^2\right) = \sum_{i=1}^{R} \left(n_i{}^2\right) + 2 \times \sum_{l<h}^{R} n_l n_h - \sum_{i=1}^{R} \left(n_i{}^2\right) = 2 \sum_{l<h}^{R} n_l n_h \tag{7}$$

Therefore, because of (5) and (7), $D(g|X)$ can be written as

$$D(g|X) = \frac{2(P-Q)}{2\sum_{l<h}^{R} n_l n_h} = \frac{P-Q}{\sum_{l<h}^{R} n_l n_h} \tag{8}$$

Metsämuuronen (2021a) showed that $D(g|X)$ can be expressed by using the directional Jonckheere-Terpstra test statistic (*JT;* Jonckheere, 1954; Terpstra, 1952). *JT* extends the directional *U* and its computational procedure to polytomous cases (e.g., Siegel & Castellan, 1988). Therefore, following the logic of Cureton (1956), the following measure can be called rank-*poly*serial correlation (ρ_{RP}):

$$\rho_{RP} = 2 \times \frac{JT_{gX}^{Obs}}{JT_{gX}^{\text{Max}}} - 1 = 2 \times \frac{JT}{\sum_{l<h}^{R} n_l n_h} - 1 = 2 \times \frac{\sum_{l<h}^{R} U_{lh}}{\sum_{l<h}^{R} n_l n_h} - 1 \tag{9}$$

(Metsämuuronen, 2022a), where JT_{gX}^{Obs} and JT_{gX}^{Max} are the observed and maximum *JT* statistics, JT_{gX}^{Obs} is the sum of all *U* statistics obtained between the subpopulations $l<h$, and $\sum_{l<h}^{R} n_l n_h$ refers to the number of all possible directional pairs of cases in the given data set. Metsämuuronen (2021a) also showed that

$$2 \times \frac{JT}{\sum_{l<h}^{R} n_l n_h} - 1 = \frac{P-Q}{\sum_{l<h}^{R} n_l n_h} = D(g|X) = \rho_{RP} \tag{10}$$

In the binary case, $l=0$, and $h=1$, $JT = U_1$, and Equation (10) reduces to

$$2 \times \frac{U_1}{n_0 n_1} - 1 = 1 - 2 \times \frac{U_0}{n_0 n_1} = D(g|X) = \rho_{RB} \tag{11}$$

Thus, Somers *D* is a rank-polyserial correlation coefficient in the polytomous setting and a rank-biserial correlation coefficient in the binary setting (Metsämuuronen, 2022a).

Since Cureton's effect size is strictly related to Somers delta, it is worth noting that the delta is also the mother of many other estimators and statistics (see Newson, 2002, 2006). These include the Gehan-Breslow test for censored outcomes (Breslow, 1970; Gehan, 1965) based on the Wilcoxon W test (Wilcoxon, 1945) and the Kruskal-Wallis test (Kruskal & Wallis, 1952), the Gini index (Gini, 1909), and Harrell's *C* index (Harrell, 2001) for censored outcomes (Harrell et al., 1982), also known as Nelson's *V* (Nelson, 1984). Because Harrell's *C* is analogous to the area under the receiver operating characteristic (ROC) curve (AUC; e.g., Heagerty & Zheng, 2005), and because Somers' *D* can be expressed in terms of *C*: $D = 2C - 1$ (Harrell, 2001), AUC is also a function of Somers' *D*: AUC = $\frac{1}{2} \times (D+1)$ (see also Jansen, 1984; Newson, 2006; Smithson, 2023).

PHD, "probability of higher subgroup dominance" based on Somers D

The element $JT/\sum_{l<h}^{R} n_l n_h$ embedded in Somers' *D* in Equation (10) is of particular interest. Its interpretation is straightforward: it strictly indicates the proportion of the ascending located observations after ordering them by *X* (see discussion in Metsämuuronen, 2021a). From a probability viewpoint, it indicates the probability of higher subgroup dominance in the sense that if we were to take 100 random pairs from the data set with different subpopulations, it indicates

the proportion of pairs where the case from the higher ranked group also scored higher in X, regardless of the number of groups. The following statistic can then be called the "proportion of cases in ascending order" or the "probability of higher subgroup dominance based on Somers' D" or, in short, the "probability of higher subgroup dominance" (*PHD*):

$$PHD = 0.5 \times D(g|X) + 0.5 \tag{12}$$

The statistic itself is discussed by Metsämuuronen (2021a, 2022a) in connection with the interpretation of Somers' D, and the interpretation of deflation-corrected reliability estimates (e.g., Metsämuuronen, 2022e, 2023c). However, it has not been associated with the effect sizes and, in particular with the common language effect sizes, in any published work.

Equation (12) seems to be a new general estimator of ES, which could also be called either "probability of superiority of the higher subgroup", an index of "stochastic superiority of the higher subgroup" or "higher subgroup dominance statistic", mimicking the epithets of *PS*, *A* and Cliff's *d*. This is because it directly indicates the dominance or superiority of the higher of the subpopulations over the lower of the subpopulations. If $D = 0.90$, $PHD = 0.5 \times 0.9 + 0.5 = 0.95$. Using the classic common language interpretation of McGraw and Wong (1992) and extending it to polytomous ordinal settings, we infer that in 95 out of 100 random pairs from different subpopulations, regardless of the number of subpopulations, the case from the higher-ranked subpopulation was superior (in terms of magnitude in the metric variable) to the case from the lower-ranked subpopulation. *PS*, *d* and *A* express the same from the opposite perspective: in 5 out of 100 random pairs, subpopulation 0 was superior to subpopulation 1 (see Appendix 1).

The *PHD* is general in two ways. First, whereas *PS, d,* and *A* are algebraically restricted to the dichotomous cases because of the interdependence of the U statistic in their basic form (see Appendix 1), Somers' D and the related *PHD* are algebraically related to both dichotomous and polytomous ordinal cases because of the interdependence of the *JT* statistic, which generalizes U to the polytomous settings. Second, in their basic form, *PS, d,* and *A* can be expressed by using the formula of *PHD*:

$$\begin{cases} PS = A = 1 - PHD \\ d = 1 - 2PHD \end{cases} \tag{13a}$$

(see Appendix 1). It may be worth emphasizing again that Equation (13a) gives forms for *PS*, *A* and *d* that can be used strictly with polytomous ordinal cases. This is discussed in Section "Three novel CLES estimators for ordinal settings based on Somers' D".

As discussed above, many well-known statistics are special cases of Somers' D. From the perspective of *PHD*, there is an interesting relationship between the effect size associated with the area under the ROC (AUC; see Smithson, 2023; Ruscio & Mullen, 2012) and *PHD*. Since AUC = $1/2 \times (D + 1)$ and $D = 2 \times PHD - 1$, consequently, AUC = *PHD*. However, this relationship will not be discussed further.

Three novel CLES estimators for ordinal settings based on Somers' D

As shown in Appendix 1, in their form relevant to binary and dichotomous settings, *PS, d,* and *A* are strictly interdependent of Somers' D. In the polytomous ordinal case, it seems justified to think that Equation (13a) (see also Equations A3, A7, and A12 in Appendix 1) leads to three new effect size estimators. This is most obvious with *A*.

Vargha and Delaney (2000) solve the polytomous case in a fundamentally different way than Equation (13a). In the Vargha and Delaney's algorithm, *A* compares a single category either with a *union* of the remaining categories or by computing *A* for all pairwise comparisons and aggregating those as *A*, and Ruscio and Gera (2013) compute *A* for the closest ordinal categories and aggregate those as *A*. *PHD* and the related *JT* and U statistics embedded in Somers' D treat all

the categories simultaneously and individually. Then, the estimates of 1–*PHD* and *A* lead to a somewhat different result (see the numerical example in Section "Estimation in the binary or dichotomous case"). Thus, it would be confusing to refer to the estimator in Equation (13a) as Vargha-Delaney *A* or Ruscio-Gera *A* in the ordinal polytomous settings. The name of the new estimator could be "ordinal *A*".

If the new estimator related to *A* is renamed (e.g., "ordinal *A*"), it may be justified to give a new name to the "ordinal *PS*" and "ordinal *d*" as well. The new estimators are summarized as follows:

$$\begin{cases} ordinal\ PS = ordinal\ A = 1 - \left(0.5 \times D\left(g|X\right) + 0.50\right) = 1 - PHD \\ ordinal\ d = -D\left(g|X\right) = 1 - 2PHD \end{cases} \tag{13b}$$

While *PHD* and the related *PHG* (see Section "PHG, "Probability of higher subgroup dominance" based on Goodman–Kruskal G") express the probability in terms of higher subgroup dominance in the same way as Somers' *D* and Goodman-Kruskal *G*, the ordinal versions of *A*, *PS,* and *d* express the probability in terms of lower subgroup dominance. In some cases, the latter logic may be better than that used in *PHD* and *PHG*. For example, consider a data set where the small rank order expresses something positive, such as in the education, a variable related to the support needed to learn. For such a variable, the base group could be coded 0 ("no need for extra support") and the other groups subsequently coded 1 ("need for extra support") and 2 ("need for intensive support"). We might be interested in knowing what the effect size of this grouping is with respect to the achievement level (or, alternatively, of how notably the means of the groups differ from each other); obviously, the worse the performance on a math test, the more support would be expected to be needed. In this case, the ordinal *A* and the ordinal *PS* logically express how many cases from the *lower* group (i.e., less support needed) score higher on the test, and the ordinal *d* expresses the same as a positive association. In this case, *PHD* express the probability in a somewhat irrelevant way.

PHG, "probability of higher subgroup dominance" based on Goodman–Kruskal G

Since *PHD* based on *D* is a general estimator of CLES in the polytomous ordinal and dichotomous settings, it is worth highlighting that Metsämuuronen (2021a, 2022a) also discusses the same formula, but uses Goodman-Kruskal gamma (*G*) instead of *D*. Both *D* and *G* estimate the probability γ that two randomly selected pairs have the same order in two variables (e.g., Metsämuuronen, 2021a; Van der Ark & Van Aert, 2015). Both belong to the same family of estimators, called either the delta family (e.g., Metsämuuronen, 2020b; Newson, 2006), the gamma family (e.g., Van der Ark & Van Aert, 2015; Woods, 2007), or the tau family (e.g., Kendall, 1948; Kendall & Gibbons, 1990). The mother of all these seems to be Kendall's tau-a (Kendall, 1938), because, if both variables are continuous, implying no tied pairs, both *D* and *G* are equal to *tau-a* (see Kendall & Gibbons, 1990; Newson, 2006).

While the related *tau-b* are truly symmetric measures, both *D* and *G* are truly directional measures even if *G* is often described as a symmetric measure (see, e.g., Sheskin, 2011; Sirkin, 2006; Wholey et al., 2015; see also the output of, e.g., IBM SPSS software; see the algebraic discussion in Metsämuuronen, 2021a). Namely, *G* is a special case of *D* although the opposite is sometimes claimed (e.g., Kvålseth, 2017): when there are no tied pairs but the variables are not continuous,

$$G = D\left(g|X\right) \tag{14}$$

In this case, *G* does not equal with the other directions of *D*, i.e.,

$$G \neq D\left(X|g\right) \neq D(Sym) \tag{15}$$

(Metsämuuronen, 2021a). Therefore, G could be expressed also as $G(g|X)$. If both g and X are continuous, $D(g|X) = D(X|g) = D(Sym) = G = Tau\text{-}a$ (see Newson, 2002) as discussed above.

From the viewpoint $G(g|X)$, $D(g|X)$ can be expressed as follows:

$$D(g|X) = \frac{P - Q}{P + Q + T} \tag{16}$$

(Metsämuuronen, 2021a, cl. Equations 5 and 8; note that in this form, P and Q are already doubled), where T refers to the number of tied pairs.[2] G can be expressed as

$$G = G(g|X) = \frac{P - Q}{P + Q} \tag{17}$$

Comparing Equations (16) and (17), it becomes clear why the estimates of D tend to be smaller in magnitude than those of G. When there are tied pairs, i.e., when $T > 0$, the estimates of $D(g|X)$ are more conservative than those of G, i.e.,

$$\hat{G} = \hat{G}(g|X) > \hat{D}(g|X) \tag{18}$$

The obvious reason for the conservative nature of D is that while D uses all pairs as the basis for probability, G uses only those pairs where the direction is known. Both ways of thinking about probability are justified. The same logic as in G of using only relevant cases as the basis for probability is also used, for example, in the sign test and the Wilcoxon signed-rank test. Therefore, although some scholars have claimed that the estimates of G are "inflated" (see, e.g., Higham & Higham, 2019; Kvålseth, 2017; Masson & Rotello, 2009), this is not factually true. This impression may arise partly from G's hidden directional nature and partly from the way in which the basis for the probability is defined (see discussion and further literature in Metsämuuronen, 2021a).

Metsämuuronen (2022a, 2022d, 2022e, 2022f) discusses the possible use of G in the formula in Equation (12), although the algebraic relationship between G and the JT statistic is more complicated compared to D. The parallel statistic to Equation (12), $0.5 \times G + 0.5$, indicates the "proportion of ascending order cases" or the "probability of higher subgroup dominance", although the estimate is somewhat more liberal compared to Equation (12) because the unspecified directions are omitted from the analysis. Thus, if $G = 0.90$, then $(0.5 \times 0.9 + 0.5) = 0.95$ and, in common language, we infer that in 95 cases out of 100 random pairs from different subpopulations, the case from the higher subpopulation was superior (in terms of magnitude in the metric variable) to the case from the lower subpopulation, *considering only those cases where the difference in the metric variable was obtained.*

In summary, we have two general effect size estimators that can be used for binary, dichotomous, and ordinal settings based on either D or G. To distinguish between their names, the one based on D is abbreviated by *PHD* and the one based on G by *PHG*, and both express the "probability of higher subgroup dominance":

$$\begin{cases} PHD = 0.5 \times D(g|X) + 0.5 = 0.5 \times D + 0.5 \\ PHG = 0.5 \times G(g|X) + 0.5 = 0.5 \times G + 0.5 \end{cases} \tag{19}$$

Adding here the three generalizations and modifications of the traditional CLES—ordinal *PS*, ordinal *d*, and ordinal *A*—we have a set of five novel CLES estimators suitable for ordinal data sets. A numerical example illustrates the differences and similarities between the estimators.

Table 1. Hypothetical data set for estimating effect sizes for ordinal settings.

g1	g2	X	X ranked by g1	X ranked by g2 between groups 0 and 1	X ranked by g2 between groups 1 and 2	X ranked by g2 between groups 0 and 2
0	0	7	1	1		1
0	0	12	2	2		2
1	1	13	3	3	1	
0	0	14	4	4		3
0	0	15	5,5	5,5		4
1	1	15	5,5	5,5	2	
0	0	16	7	7		5
0	0	17	8,5	8,5		6
1	1	17	8,5	8,5	3	
0	1	18	10	10	4	
0	0	20	11	11		7
0	1	21	12,5	12	5,5	
1	2	21	12,5		5,5	8
1	1	25	14	13	7	
0	0	29	15	14		9
1	2	31	16		8	10
1	2	32	17		9	11
1	2	33	18		10	12
0	2	34	19,5		11,5	13
1	1	34	19,5	15	11,5	
0	1	37	21,5	16	13,5	
1	2	37	21,5		13,5	14
1	2	38	23		15	15

Numerical examples of the ordinal CLESs

Hypothetical dataset

Consider a hypothetical data set as shown in Table 1 with two types of grouping variables (g1 and g2). Of these, g1 is a binary one with $n_0 = 12$ and $n_1 = 11$. The subpopulations could be related to gender (or sex) or dummy variables or experimental design where 1 refers to the experimental or "favorable" group or the group of interest. The polytomous variable g2 with $n_0 = 8$, $n_1 = 8$, and $n_2 = 7$ can refer to an ordinal categorical variable (such as the level of education, e.g., primary, secondary, and tertiary, or three levels of socioeconomic status) or an experimental design with a treatment with an incremental effect (such as a series of lectures, e.g., 10, 15, or 20 h or a dose of medication of, e.g., 0.05, 0.10, 0.15 mg). The data set is ordered by the score variable (X), and the different sets of rankings of the score are computed ("X ranked by … ").

The manual calculation of the estimates is illustrated in an Excel file found at https://doi.org/10.13140/RG.2.2.10199.18089, and a designated R package *CLES* (Niemensivu & Metsämuuronen, 2024) may be useful for R users at https://github.com/TNiemensivu/CLES.

Estimation of D and G with the traditional tools

In traditional software packages such as IBM SPSS, the syntax for D is `CROSSTABS/TABLES=item BY Score/STATISTICS=D` and the syntax for G is `CROSSTABS/TABLES=item BY Score/STATISTICS=GAMMA`. In STATA, the command `tabulate g X [if] [in] [weight] [, gamma]` produces G (see Stata Corp, 2018), and Newson's module (1998) produces D. In SAS, the command PROC FREQ provides G and D by specifying the TEST statement with the GAMMA, D, SMDCR options. Similarly, by using the R package *DescTools*, for example, D can be computed by `SomersDelta(x, y=NULL, direction=c("row", "column"), conf.level=NA, …)` and G by `Goodman`

`KruskalGamma(x, y=NULL, conf.level=NA, ...)` (see https://rdrr.io/cran/DescTools/man/). In the output related to *D*, select the "*X* dependent" option. The R package *CLES* (https://github.com/TNiemensivu/CLES) computes *G* and strictly the appropriate direction for *D*. The syntax that gives also the estimates of *D* and *G* is `>summary(all([variable with the narrower scale],[variable with the wider scale]))`, e.g., for variable *g*1 and *X* as `>summary(all(g1,X))`.

Once the estimates of *D* and *G* have been obtained, *PHD* and *PHG* (as well as the ordinal *PS*, ordinal *d*, and ordinal *A*) can be calculated by using, for example, a common spreadsheet program (see Excel template at https://doi.org/10.13140/RG.2.2.10199.18089). The R package *CLES* by Niemensivu and Metsämuuronen (2024) computes the estimates by the commands `>summary(all(g1,X))` and `>summary(all(g2,X))`.

Estimation in the binary or dichotomous case

Calculating the estimates for the binary case is straightforward. By adding the rankings of the higher group 1 from the column "*X* ranked by *g*1", the Wilcoxon test statistic for group 1 is as follows: $W_1 = R_1 = (3 + 5.5 + 8.5 + \cdots + 21.5 + 23) = 158.5$ and $n_1(n_1 + 1)/2 = 11 \times 12/2 = 66$. Then, the resulting Mann–Whitney U_1 is $U_1 = R_1 - n_1(n_1 + 1)/2 = 158.5 - 66 = 92.5$ and $PHD = \frac{U_1}{n_0 n_1} = \frac{92.5}{12 \times 11} = 0.701$.

Alternatively, for the more general use, the *PHD* can be calculated using Somers's *D*. For the manual calculation of *D* and *G*, the data set regarding g1 is transformed into a contingency table (Table 2).

For *D* and *G*, two intermediate statistics are formed: C_{ij} and D_{ij} which refer to the concordant and discordant pairs between the variables. In common language, C_{ij} is computed by calculating the number of cases in each cell to the "north-west" (C_{ij} left) and "south-east" (C_{ij} right) and D_{ij} to the "south-west" (D_{ij} left) and "north-east" (D_{ij} right) in each cell. Technically, $C_{ij} = \sum_{h<i}\sum_{k<j} n_{hk} + \sum_{h>i}\sum_{k>j} n_{hk}$ and $D_{ij} = \sum_{h<i}\sum_{k>j} n_{hk} + \sum_{h>i}\sum_{k<j} n_{hk}$ (see later Equation 20 where the asymptotic standard errors of *PHD* and *PHG* are derived). However, as is a common practice in manual calculations, only *half* of the values (usually "south-east" or "south-west") are computed and doubled. The number of concordant pairs is $P = \sum_{i,j} n_{ij}C_{ij}$ and the number of discordant pairs is $Q = \sum_{i,j} n_{ij}D_{ij}$. Then, calculating only one way, $P = \sum_{i,j} n_{ij}C_{ij} = 11 + 11 + 0 + \cdots + 0 + 2 + 1 = 90$ and $Q = \sum_{i,j} n_{ij}D_{ij} = 0 + 0 + 1 + = 0 + 0 + 1 + \ldots + 8 + 9 + 0 = 37$. Using these symbols and Equation (5), and doubling *P* and *Q*, the estimate of *D* is $D(g|X) = \frac{2 \times (90-37)}{529^2 - (144+121)} = 0.402$ and by using Equation (17), the estimate of *G* is $G = \frac{90-37}{90+37} = 0.417$. Using these rank correlation estimates, $PHD = 0.5 \times D + 0.5 = 0.5 \times 0.402 + 0.5 = 0.701$ and $PHG = 0.5 \times G + 0.5 = 0.5 \times 0.417 + 0.5 = 0.709$.

PHD strictly indicates that 70.1% of the pairs in the obtained data set are such that the case from the higher-ranked subgroup (1) scores higher on the metric variable than the case from the lower-ranked subgroup (0). Alternatively, out of (the hypothetical) 1000 random pairs, 701 cases from group 1 are expected to be superior to the cases from group 0. Accordingly, *PHG* indicates that 70.9% of the pairs in the obtained data set are such that the cases from the higher-ranked subgroup score higher on the metric variable than the case from the lower-ranked subgroup, if

Table 2. Contingency table based on Table 1 (g1 vs. X).

		X																			
		7	12	13	14	15	16	17	18	20	21	25	29	31	32	33	34	37	38	Total	n_i^2
g1	0	1	1	0	1	1	1	1	1	1	1	0	1	0	0	0	1	1	0	12	144
	1	0	0	1	0	1	0	1	0	0	1	1	0	1	1	1	1	1	1	11	121
Total		1	1	1	1	2	1	2	1	1	2	1	1	1	1	1	2	2	1	23	529

only those cases are considered where cases have a different value on the metric variable. Alternatively, out of (the hypothetical) 1000 random pairs in the population, 709 cases from group 1 are expected to be superior to the cases from group 0 when only the relevant pairs are considered.

For reference, the estimates of the traditional CLES estimators are also computed. Using Ruscio's *RProbSup* package (2022), the traditional Vargha-Delaney *A* estimate for the binary variable g1 is $A_{g1} = 0.701$, which is identical to the *PHD* estimate.[3] Logically, the estimate should be $A_{g1} = 0.299$ (= 1–*PHD*) because *A* estimates the superiority of the first group (0) over the second group (1) in the same way as *PS* (see Equation 13a; see Appendix 1). Similarly, because of Equation (13a), the traditional Grissom-Kim *PS* gives the value $PS_{g1} = 0.299$ (= 1–*PHD*) and Cliff's *d* gives the value $d_{g1} = -0.402$ (= 1–2*PHD* = –*D*). Of these, the interpretation of $A = PS = 0.299$ gives the strict probability that cases from subpopulation 0 are superior to cases from subpopulation 1 with respect to the metric variable *X*: of all pairs, 29.9% were such that the case from the category 0 scored higher in *X*. $PHD = 0.701$ gives the probability that the subpopulation 1 is superior to the subpopulation 0: of all pairs, 70.1% were such that the case from the category 1 scored higher in *X*.

Estimation of a polytomous ordinal case

From the point of view of the new estimators, the polytomous ordinal variable *g*2 may be more interesting than the binary g1. The statistics needed for the calculation can be computed by using *JT* statistics or strictly by computing *D* and *G*. Let us start with *JT* which leads to a parallel notation as above with the binary settings. Applying Equations (9) and (10) to the three groups 0, 1 and 2, we get

$$D(g|X) = 2 \times \frac{JT}{\sum_{i<j}^{R} n_i n_j} - 1 = 2 \times \frac{U_{01} + U_{12} + U_{02}}{n_0 n_1 + n_1 n_2 + n_0 n_2} - 1$$
$$= 2 \times \frac{\left(R_{01} - \frac{n_1(n_1+1)}{2}\right) + \left(R_{12} - \frac{n_2(n_2+1)}{2}\right) + \left(R_{02} - \frac{n_2(n_2+1)}{2}\right)}{n_0 n_1 + n_1 n_2 + n_0 n_2} - 1$$

where R_{01} is the sum of the ranks in *X* associated with the higher of the groups 0 and 1, R_{12} is the sum of the ranks in *X* associated with the higher of the groups 1 and 2, and R_{02} is the sum of ranks in *X* associated with the higher of the groups 0 and 2. These ranks are given in Table 1: $R_{01} = 3 + 5.5 + \cdots + 15 + 16 = 83$, $R_{12} = 5.5 + 8 + \cdots + 13.5 + 15 = 72.5$, and $R_{02} = 8 + 10 + \cdots + 14 + 14 = 83$. Then, $D(g|X) = 2 \times \frac{(83 - \frac{8(8+1)}{2}) + (72.5 - \frac{7(7+1)}{2}) + (83 - \frac{7(7+1)}{2})}{8\times8 + 8\times7 + 8\times7} - 1 = 2 \times \frac{146.5}{176} - 1 = 0.665$ and, therefore, $PHD = 0.5 \times 0.665 + 0.5 = 0.832$. This strictly indicates a condition where in 83.2% of all the pairs from different categories in *g* that can be formed by the data set obtained, the case from a higher-ranked subgroup will score higher on *X* than the case from a lower-ranked subgroup regardless of which of the three categories the cases come from. Alternatively, out of (the hypothetical) 1000 random pairs in population, we expect to see 832 cases from a higher-ranked subpopulation to score higher on *X* than the case from the lower-ranked subpopulation.

As a reference for the polytomous variable g2, Vargha and Delaney (2000) have proposed two variants of *A* and Ruscio and Gera (2013) discussed of two more, one of which is relevant here. [4] Here, the ordinal variant, A_{ord}, provided by Ruscio and Gera (2013) is discussed, because it comes the closest to the interpretation of the directional *D*, *PHD* and ordinal *A*. In this method, each individual ordinal category is compared to the adjacent group after which the mean of these values is the estimate of *A*. This estimate of *A* for the polytomous item g2 is $A_{ord} = 0.235$.[5] Using Equation (13b), *ordinal* A_{g2} = *ordinal* PS_{g2} = 1–*PHD* = 1–0.832 = 0.168. The ordinal version of Cliff's *d* gives the estimate *ordinal* $d_{g2} = -D = -0.665$. Note that Ruscio and Gera's procedure does not yield identical estimates to *PHD,* and the difference is notable; *ordinal A* indicates higher effect size than A_{ord}. The reason for the different result lies in the way the subpopulations

Table 3. Contingency table based on Table 1 (g2 vs. X).

		X																			
		7	12	13	14	15	16	17	18	20	21	25	29	31	32	33	34	37	38	Total	n_i^2
g2	0	1	1	0	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	8	64
	1	0	0	1	0	1	0	1	1	0	1	1	0	0	0	0	1	1	0	8	64
	2	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	7	49
Total		1	1	1	1	2	1	2	1	1	2	1	1	1	1	1	2	2	1	23	529

are treated. Whereas *PHD* and ordinal *A* and the embedded *JT* and *U* statistics treat all the categories simultaneously and individually, Ruscio and Gera's *A* compares individual subpopulations with the closest ordinal subpopulation, and aggregates the resulting individual *A* statistics as *A*. Thus, in the polytomous ordinal settings, the results obtained by *PHD* and the associated *ordinal A* are more specific than the A_{ord} of Ruscio and Gera.

While the interpretation of the *PHD* in the polytomous ordinal case is unambiguous because the interpretation of the *JT* and *U* test statistics is unambiguous (0.832 refers strictly to the theoretical 832 superior cases out of 1000, or 83.2% of the pairs), the interpretation of Ruscio and Gera's polytomous estimate $A_{ord} = 0.235$ (or 0.765) is more complicated, and it does not fully correspond to McGraw and Wong (1992) original idea of the common language effect size estimate. While the *ordinal* $A = 0.168$ has a strict common language interpretation ("168 pairs out of 1000"), we do not know what an estimate of $A = 0.235$ means. It should be noted, however, that because of the algorithm, the other two traditional extensions of *A* that ask more general questions, A_{AAD} and A_{AAPD}, give more relevant estimates with polytomous *nominal* settings where *PHD* and *PHG* do not make sense, nor do the *ordinal A*, *PS*, and *d*. Moreover, *PHD*, *PHG*, and the *ordinal A* are not appropriate in the correlated data sets where the other variants of *A* are relevant (see Ruscio & Gera, 2013; Vargha & Delaney, 2000).

PHG could be calculated by using the *JT* statistic to calculate *G* for *PHG* as above with *D* and *PHD*: $G = 2 \times \frac{JT}{(\sum_{l<h}^{R} n_l n_h) - T} - 1 = 2 \times \frac{JT}{P+Q} - 1$ (Metsämuuronen, 2021a). However, the following illustrates the traditional way of calculating *G* (and *D*) using the contingency table between *g2* and *X* (Table 3).

Now, as calculated one way, $P = \sum_{i,j} n_{ij} C_{ij} = 15 + 15 + 0 + \cdots + 0 + 2 + 1 = 144$ and $Q = \sum_{i,j} n_{ij} D_{ij} = 0 + 0 + 1 + \cdots + 4 + 5 + 0 = 27$ (see the real calculation at https://doi.org/10.13140/RG.2.2.10199.18089). Using these symbols and the sample sizes from Table 3, the estimates of *D* and *G* are $D(g|X) = \frac{2 \times (144-27)}{529^2 - (64+64+49)} = 0.665$ and $G = \frac{144-27}{144+27} = 0.684$. Consequently, $PHD = 0.5 \times 0.665 + 0.5 = 0.832$ and $PHG = 0.5 \times 0.684 + 0.5 = 0.842$.

For *PHD*, the interpretation is the same as above: out of (the hypothetical) 1000 random pairs in the population, we expect to see 832 such pairs (83.2%) where a case from a higher-ranked subpopulation scores higher on *X* than the case from the lower-ranked subpopulation. Accordingly, *PHG* indicates that the percentage would be 84.2% if only the cases where the superiority is known were considered. Confidence intervals and statistical significance of the estimates are discussed below.

Asymptotic standard errors of the ordinal CLESs

The elegance of the estimators in Equations (19) and (13b) is that, because the asymptotic standard errors (ASEs) of *D* and *G* are known (e.g., Agresti, 2010; Goodman & Kruskal, 1979; IBM, 2024; Metsämuuronen, 2021a, 2021b; Somers, 1980), the ASEs of all five estimators discussed above are known. These are derived below.

Let us define the following statistics:

$$\begin{aligned} C_{ij} &= \sum_{h<i}\sum_{k<j} n_{hk} + \sum_{h>i}\sum_{k>j} n_{hk} \\ D_{ij} &= \sum_{h<i}\sum_{k>j} n_{hk} + \sum_{h>i}\sum_{k<j} n_{hk} \\ P &= \textstyle\sum_{i,j} n_{ij} C_{ij} \\ Q &= \textstyle\sum_{i,j} n_{ij} D_{ij} \\ D_g &= N^2 - \sum_{j=1}^{R} \left(n_i{}^2\right), \end{aligned} \tag{20}$$

where n_{ij} is the number of cases in the cell ij of the two-way contingency table. In the following, the ASEs based on the normal approximation are discussed first in Section "ASEs based on the normal approximation", followed by the alternatives based on the uniform population in Section "ASEs based on the uniform approximation". Of these, the latter may be more useful for samples where, for example, all cases have a unique score. A numerical example of ASEs based on Table 1 is given in Section "Numeric example of ASEs".

ASEs based on the normal approximation

Assuming that the observations are sampled from a normally distributed population, the sampling variance for $D(g|X)$ for the calculation of confidence intervals is as follows:

$$VAR_1\left(D\left(g|X\right)\right) = \frac{4}{D_g^4}\left(\sum_{i,j} n_{ij}\left[D_g\left(C_{ij} - D_{ij}\right) - (P-Q)(N-n_i)\right]^2\right) \tag{21}$$

Then, using the standard formula for variance, the sampling variance for *PHD* to calculate confidential intervals is

$$\begin{aligned} VAR_1(PHD) &= VAR_1\left(0.5D\left(g|X\right) + 0.5\right) = 0.5^2 \times VAR\left(D\left(g|X\right)\right) \\ &= \frac{1}{D_g^4}\left(\sum_{i,j} n_{ij}\left[D_g\left(C_{ij} - D_{ij}\right) - (P-Q)(N-n_i)\right]^2\right) \end{aligned} \tag{22}$$

Consequently, the ASE for *PHD* for calculating confidence intervals can then be computed as follows:

$$\begin{aligned} ASE_1(PHD) &= \sqrt{VAR_1(PHD)} \\ &= \frac{1}{D_g^2}\sqrt{\sum_{i,j} n_{ij}\left[D_g\left(C_{ij} - D_{ij}\right) - (P-Q)(N-n_i)\right]^2} \end{aligned} \tag{23a}$$

Alternatively,

$$ASE_1(PHD) = 0.5 \times ASE_1\left(D\left(g|X\right)\right) \tag{23b}$$

Since the *ordinal PS* = *ordinal A* = 1–*PHD*, consequently, because of VAR_1(ordinal *PS*) = VAR_1(ordinal *A*) = $VAR_1(PHD)$ = 0.25×$VAR_1(D(g|X))$. Then, ASEs for the *ordinal PS* and *ordinal A* for calculation of the confidence intervals can be computed as follows:

$$\begin{aligned} ASE_1(ordinal\ PS) &= ASE_1(ordinal\ A) \\ &= \frac{1}{D_g^2}\sqrt{\sum_{i,j} n_{ij}\left[D_g\left(C_{ij} - D_{ij}\right) - (P-Q)(N-n_i)\right]^2} \end{aligned} \tag{24}$$

Since the *ordinal d* is equal to –*D*, $VAR_1(ordinal\ d) = VAR_1(D(g|X))$. Then, ASE for the *ordinal d* for the calculation of confidence intervals can then be computed as follows:

$$ASE_1(ordinal\ d) = \frac{4}{D_g^2}\sqrt{\sum_{i,j} n_{ij}\left[D_g(C_{ij} - D_{ij}) - (P - Q)(N - n_i)\right]^2} \tag{25}$$

The sampling variance for $D(g|X)$ under the hypothesis of independence (i.e., that $D = 0$) is as follows

$$VAR_0\big(D\big(g|X\big)\big) = \frac{4}{D_g^2}\left(\sum_{i,j} n_{ij}(C_{ij} - D_{ij})^2 - \frac{1}{N}(P - Q)^2\right) \tag{26}$$

and the corresponding sampling variance for *PHD* is

$$VAR_0(PHD) = \frac{1}{D_g^2}\left(\sum_{i,j} n_{ij}(C_{ij} - D_{ij})^2 - \frac{1}{N}(P - Q)^2\right) \tag{27}$$

Consequently, the ASE for *PHD* for testing the null hypothesis that $PHD = 0.5$ is

$$ASE_0(PHD) = \frac{1}{D_g}\sqrt{\sum_{i,j} n_{ij}(C_{ij} - D_{ij})^2 - \frac{1}{N}(P - Q)^2} \tag{28a}$$

or, alternatively,

$$ASE_0(PHD) = 0.5 \times ASE_0\big(D\big(g|X\big)\big) \tag{28b}$$

Given the strict relationship with *ordinal PS*, *ordinal A*, and *PHD*, the ASEs for the *ordinal PS* and *ordinal A* for for testing the null hypothesis that *ordinal PS* = *ordinal A* = 0.5 can be computed as follows:

$$\begin{aligned} ASE_0(ordinal\ PS) &= ASE_0(ordinal\ A) \\ &= \frac{1}{D_g}\sqrt{\sum_{i,j} n_{ij}(C_{ij} - D_{ij})^2 - \frac{1}{N}(P - Q)^2} \end{aligned} \tag{29}$$

Parallel to the ASE1 for *ordinal d*, also the ASE0 for testing that *ordinal d* = 0, is equal to the ASE0 of *D*:

$$ASE_0(ordinal\ d) = \frac{4}{D_g}\sqrt{\sum_{i,j} n_{ij}(C_{ij} - D_{ij})^2 - \frac{1}{N}(P - Q)^2} \tag{30}$$

The ASE for *G* which is used to estimate the confidence interval, it can be computed as follows:

$$ASE_1(G) = \frac{4}{(P + Q)^2}\sqrt{\sum_{i,j} n_{ij}(QC_{ij} - PD_{ij})^2} \tag{31}$$

and, therefore, for *PHG,* as follows:

$$ASE_1(PHG) = \frac{2}{(P + Q)^2}\sqrt{\sum_{i,j} n_{ij}(QC_{ij} - PD_{ij})^2} \tag{32a}$$

or

$$ASE_1(PHG) = 0.5 \times ASE_1(G) \tag{32b}$$

Finally, the ASE for G assuming independence ($G = 0$), is as follows:

$$ASE_0(G) = \frac{2}{(P+Q)}\sqrt{\sum_{i,j} n_{ij}(C_{ij} - D_{ij})^2 - \frac{1}{N}(P-Q)^2} \tag{33}$$

and, consequently, for PHG to test that $PHG = 0.5$ is

$$ASE_0(PHG) = \frac{1}{(P+Q)}\sqrt{\sum_{i,j} n_{ij}(C_{ij} - D_{ij})^2 - \frac{1}{N}(P-Q)^2} \tag{34a}$$

or

$$ASE_0(PHG) = 0.5 \times ASE_0(G) \tag{34b}$$

ASEs based on the uniform approximation

Because D and G are easy to use and applicable to small sample sizes (e.g., 20–30 cases) and settings with wide X leading to situation where all or almost all cases a related to an individual value of X, we may not want to assume that the population is normally distributed over the cells as above. Somers (1980; see also Siegel & Castellan, 1988; Metsämuuronen, 2020b) provides us with alternative options for the sampling variances by assuming that the observations are sampled from the *uniformly* distributed population. This may be a reasonable assumption for the cases where we simultaneously have a small sample and a large (or continuous) scale in the metric variable; indeed, it may be the case that all cases receive a unique value in the metric variable even in the population. For example, we face this condition in the numerical example in Table 1 above. Also, in many classroom tests with small populations, the uniform approximation may be more relevant than the normal approximation.

Under this assumption, the sampling variance for testing the null hypothesis for $D(g|X)$ and ordinal d is given as:

$$VAR_0\big(D\big(g|X\big)\big) = VAR_0(ordinal\ d) = \frac{4(C^2-1)(R+1)}{9NC^2(R-1)} \tag{35}$$

and for gamma as follows:

$$VAR_0(G) = \frac{4(C+1)(R+1)}{9N(C-1)(R-1)} \tag{36}$$

(Somers, 1980), where C refers to the number of categories in the metric variable and R to the number of categories in the categorical, ordinal variable, and N is the total number of cases. Thus, assuming a uniform distribution of the metric variable, the ASE for D and PHD for statistical testing are calculated as follows:

$$ASE_0(D) = \sqrt{\frac{4(C^2-1)(R+1)}{9NC^2(R-1)}} \tag{37a}$$

$$ASE_0(PHD) = 0.5 \times \sqrt{\frac{4(C^2-1)(R+1)}{9NC^2(R-1)}} = \sqrt{\frac{(C^2-1)(R+1)}{9NC^2(R-1)}} \tag{37b}$$

and for G and PHG as follows:

$$ASE_0(G) = \sqrt{\frac{4(C+1)(R+1)}{9N(C-1)(R-1)}} \tag{38a}$$

$$ASE_0(PHG) = 0.5 \times \sqrt{\frac{4(C+1)(R+1)}{9N(C-1)(R-1)}} = \sqrt{\frac{(C+1)(R+1)}{9N(C-1)(R-1)}} \tag{38b}$$

The parallel ASE_0 for *ordinal PS* and *ordinal A* for testing the probability 0.5 is as follows:

$$ASE_0(ordinal\ PS) = ASE_0(ordinal\ A) = \sqrt{\frac{(C^2-1)(R+1)}{9NC^2(R-1)}} \tag{39}$$

and for *ordinal d*, the same as for *D*:

$$ASE_0(ordinal\ d) = \sqrt{\frac{4(C^2-1)(R+1)}{9NC^2(R-1)}} \tag{40}$$

Metsämuuronen (2020b) notes that, when assuming that the observations were sampled from a uniformly distributed population, the approximation of the sampling variance depends only on the number of cases, and the number of categories and rows in the variables of interest. This information may be available in the published studies pooled for meta-analyses.

Numeric example of ASEs

As a numerical example, the binary variable g1 in Table 1 is reanalyzed. The variable g2 is not analyzed here in order to condense the text. However, the Excel template at https://doi.org/10.13140/RG.2.2.10199.18089 illustrates the manual calculation of g2. After applying Equations (23b) (24). (25), (28b), (29), (30), (32b), and (34b), we obtain the ASE as follows assuming normal population (see the Excel template at https://doi.org/10.13140/RG.2.2.10199.18089 for details):

$$\begin{aligned} ASE_1(PHD_{g1}) &= ASE_1\left(ord\ PS_{g1}\right) = ASE_1\left(ord\ A_{g1}\right) \\ &= 0.5 \times \frac{2}{\left(23^2 - 11^2 - 12^2\right)^2} \times 7644.854 = 0.10969 \end{aligned}$$

$$\begin{aligned} ASE_0(PHD_{g1}) &= ASE_0\left(ord\ PS_{g1}\right) = ASE_0\left(ord\ A_{g1}\right) \\ &= 0.5 \times \frac{2}{23^2 - 11^2 - 12^2} \times 28.9738 = 0.10975 \end{aligned}$$

$$ASE_1\left(ord\ d_{g1}\right) = \frac{2}{\left(23^2 - 11^2 - 12^2\right)^2} \times 7644.854 = 0.2194$$

$$ASE_0\left(ord\ d_{g1}\right) = \frac{2}{23^2 - 11^2 - 12^2} \times 28.9738 = 0.2195$$

$$ASE_1(PHG_{g1}) = 0.5 \times \frac{4}{(180+74)^2} \times 3637.156 = 0.1128$$

$$ASE_0(PHG_{g1}) = 0.5 \times \frac{2}{(180+74)} \times 28.9736 = 0.1141$$

Based on the normal approximation, the 95% confidence intervals for the *PHD* and *PHG* estimates are $0.653 < PHD_{g1} < 0.748$ and $0.660 < PHG_{g1} < 0.757$. For *ordinal PS* and *ordinal A*, the interval is $0.252 < ordinal\ PS_{g1} = ordinal\ A_{g1} < 0.347$. When testing whether the estimate is equal to 0.5 (i.e., whether *D* or *G* is equal to zero), $Z_{PHDg1} = (0.701 - 0.5)/0.1097 = 1.829$ and $Z_{PHGg1} = (0.709 - 0.5)/0.1141 = 1.829$. For the *ordinal PS* and *ordinal A*, the test statistic is $Z_{ord\ PSg1} = Z_{ord\ Ag1} = (0.299 - 0.5)/0.1097 = -1.829$. We find that the results are identical from a statistical inference point of view: based on the standardized normal distribution, the estimates are not convincingly statistically significantly different from 0.5 (2-tailed $p = 2 \times 0.0337 = 0.067$).

Alternatively, assuming a uniform population, the ASE for the null hypotheses for *PHD* is $ASE_0(PHD_{g1}) = 0.5 \times \sqrt{\frac{4\times(18^2-1)(2+1)}{9\times23\times18^2\times(2-1)}} = 0.1202$ and the ASE for *PHG* is $ASE_0(PHG_{g1}) = 0.5 \times \sqrt{\frac{4(18+1)(2+1)}{9\times23(18-1)(2-1)}} = 0.1273$. The ASE_0 for *ordinal PS* and *ordinal A* is equal to the ASE_0 for *PHD*. Note that the estimates are slightly larger than when the normal approximation is used, and that the estimates for *PHD* and *PHG* are lightly different. This approximation yields to $Z_{PHDg1} = (0.701 - 0.5)/0.1202 = 1.670$ for *PHD* (2-tailed $p = 0.094$), to $Z_{PHGg1} = (0.709 - 0.5)/0.1273 = 1.6395$ for *PHG* (2-tailed $p = 0.101$), $Z_{ord\ PSg1} = Z_{ord\ Ag1} = (0.299 - 0.5)/0.1202 = -1.670$ for *ordinal PS* and *ordinal A* (2-tailed $p = 0.094$), and to $Z_{ord\ dg1} = (-0.402 - 0)/0.2545 = -1.670$ for *ordinal d* (2-tailed $p = 0.094$), and all refer to an estimate that is not statistically significantly different from 0.5 (or 0 in the case of *ordinal d*). Note that the z-test uses a normal approximation, although the ASE assumes uniform distribution.

Further discussion related to ASEs

Three points are worth noting of ASEs. First, unfortunately, Somers (1980, p. 115) reminds us that "although this remarkable stability of the asymptotic variances [with respect to the null hypothesis] extends to many non-null populations, the use of the simple formulas to approximate confidence intervals is not recommended". The development of possible correction factors for the ASE formulae for the violations of the assumptions caused by small samples is not the focus of this article. Systematic studies of the possible correction mechanisms would be beneficial. However, for the meta-analysis purposes, some standard options for obtaining empirical confidence intervals for *PHD* and *PHG* could be bootstrapping and jackknife techniques.

Second, even though the ASE can be calculated by using the uniform approximation and it may be closer to the true value, the statistical inference is traditionally done by using the z-test which uses the normal approximation for the probabilities. That is, even if we would obtain a test statistic $T = z = 1.67$ based on the uniform assumption, the probability for the standard value $+1.67$ is traditionally taken from the standardized normal distribution. The challenge with using the uniform distribution is that the probabilities associated with the uniform distribution depend on the minimum and maximum values of the scale.

Third, for small sample sizes, sparsely distributed contingency tables, and (possibly) non-normal populations, both the normal and uniform approximations can yield radically biased estimates for statistical significance and confidence intervals. In these situations, the most accurate estimates for the statistical significance are obtained using exact estimates (see, e.g., Mehta & Patel, 1983, 2013). In fact, Mehta and Patel (2013, p. 14) recommend computing exact probabilities "whenever the data set is small, sparse, unbalanced, or heavily tied" because the asymptotic probabilities may be far from the true value (see examples in Metsämuuronen, 2024c). For large samples, computing exact probabilities can be tedious even for the computer because it may require billions of calculations to find the "more extreme tables" used in the algorithm. If the computation takes unnecessarily long time, some algorithms, such as Mehta and Patel (2013), use Monte Carlo simulations to approximate the exact probability. Note that using the traditional Fisher-Freeman-Halton test (Fisher, 1925; Freeman & Halton, 1951) for the exact probabilities with *D* and *G* and related CLESs may lead to an opposite statistical inference than if using the algorithm of Mehta and Patel (1983, 1986, 2013). The algorithm by Mehta and Patel leads to a proper inference (see discussion in Metsämuuronen, 2024c).

When it comes to testing the null hypotheses in relation to the numerical example, it may be worth noting that the exact statistic for g1 is also known: p (exact) = 0.107 for *PHD* and p (exact) = 0.096 for *PHG*.[6] In this case, the estimates based on a normally distributed population are further from the exact value (0.029–0.040 probability units), while the estimates based on a

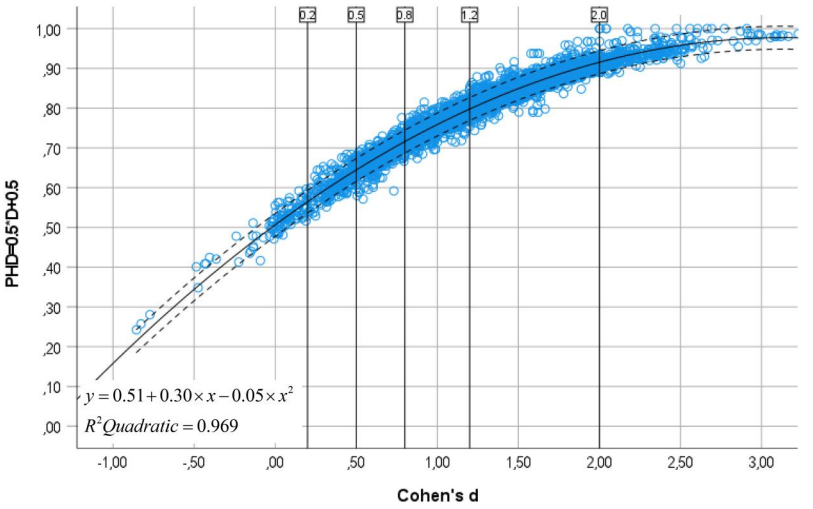

Figure 2. Relation of PHD and Cohen's d in empirical dataset (n = 7,948 estimates).

uniform population are closer (0.005–0.012 probability units). Obviously, the results cannot be generalized. However, it gives an indication that when using the uniform assumption for a small sample with a metric scale in *X*, the statistical inference may be closer to the true value. This can be valuable in meta-analyses where small samples are combined. The simplified formula is easy to use when the number of cases and the number of categories is known. Systematic studies of this phenomenon would be beneficial. In any case, the exact significance test would be recommended with small samples and sparsely distributed contingency tables with (possibly) a non-normally distributed population as discussed above.

Empirical relation of *PHD* and *PHG* with Cohen's *d*

The traditional qualitative epithets for effect size ("small", "medium", "large"), initiated by Cohen (1969, 1988), are based on the percentage of overlapping cases in two groups. These have been extended by Sawilowsky (2009) to include "very small", "small", "medium", "large", "very large", and "huge". The traditional numerical thresholds for these in the scale of Cohen's *d* are 0.1, 0.2, 0.5, 0.8, 1.2, and 2.0, respectively.

The empirical dataset discussed by Metsämuuronen (2022d) and later extended (2023a, 2023b) is published, and this makes it possible to give an empirical suggestion for the boundaries for *PHD* and *PHG* in terms of Cohen's *d*. The dataset of individual items ($n = 14{,}880$) from 1,440 tests including several indicators of the association between test items and a score variable is available in CSV format at https://doi.org/10.13140/RG.2.2.10530.76482 and in IBM SPSS format at https://doi.org/10.13140/RG.2.2.17594.72641. The procedure and characteristics of the dataset are discussed in detail by Metsämuuronen (e.g., 2022d). From this dataset, which consists of both binary and polytomous items, only the binary items ($n = 7{,}948$) are used in the following (see the discussion of the polytomous settings in Metsämuuronen, 2024a).

The dataset includes estimates of *D* and *G*, so it is easy to calculate also *PHD* and *PHG* as well. The extended dataset also includes estimates of Cohen's *d*, so it is easy to compare the estimates. Notably, the data set is somewhat specific and related to item analysis. Since the item-score correlations in published tests are usually quite high, the estimates of *D* and *G* are also quite high. However, the estimates of Cohen's *d* are also high, so the estimates are comparable. Because the estimates are on average quite high, there are only a few cases of very small or negative values of *D* and *G* in the data set. The symmetry of *D* and *G* with respect to Cohen's *d*, even

Table 4. Thresholds for effect size estimates in binary and dichotomous settings.[7]

verbal description	Cohen's d[1]	Vargha–Delaney A[2]	PHD[3,4]	PHG[3,4]	Somers D[3,5]	Goodman–Kruskal G[3]
"very small"	0.1		0.54 (0.46)	0.54 (0.46)	±0.07	±0.08
"small"	0.2	0.56 (0.44)	0.57 (0.43)	0.57 (0.43)	±0.13	±0.14
"medium"	0.5	0.64 (0.36)	0.65 (0.35)	0.65 (0.35)	±0.29	±0.31
"large"	0.8	0.71 (0.29)	0.72 (0.28)	0.73 (0.27)	±0.43	±0.45
"very large"	1.2		0.80 (0.20)	0.81 (0.19)	±0.59	±0.62
"huge"	2		0.91 (0.09)[6]	0.93 (0.07)	±0.81	±0.84

[1)] Cohen (1969, 1988); Sawilowsky (2009).
[2)] Vargha and Delaney (2000); because $A = PS$ (see Appendix 1), the same limits apply also for Grissom–Kim *PS*.
[3)] Based on the empirical data set of Metsämuuronen (2022d, 2023a, 2023b) and a quadratic model.
[4)] The same thresholds are valid also with *ordinal PS* and *ordinal A*.
[5)] The same thresholds are valid also for *ordinal* d.
[6)] $PHD = 0.91$ refers to 91 pairs out of 100, and $PHD = 0.09$ refers to 9 pairs out of 100. Both are equally "huge" in terms of effect size.

at the negative end of the correlation scale makes it possible to extend the results also to the negative side of the scale. Figure 2 illustrates the relationship between *PHD* and Cohen's *d* in the data set restricted to $d \leq 3.0$.

Using the quadratic model $PHD = 0.51 + 0.30 \times d - 0.05 \times d^2$ we can compute the limits for *PHD* corresponding to the limits given for Cohen's *d*.[8] Similarly, the model $PHG = 0.51 + 0.31 \times d - 0.05 \times d^2$ gives the limits for *PHG*. The thresholds for *D* and *G* are also computed; the model for *D* is $0.01 + 0.60 \times d - 0.1 \times d^2$ and the model for *G* is $0.02 + 0.63 \times d - 0.11 \times d^2$. The proposal based on the empirical data set are summarized in Table 4.

Four points should be made about Table 4. First, the thresholds shown for Cohen's *d*, *D*, and *G* are absolute values; the negative value indicates that the cases in the lower group(s) were superior to the cases in the higher group(s). Second, the thresholds for *A*, *PHD*, and *PHG* are always non-negative as probabilities. Because of symmetry, the low probabilities must also be considered relevant from an effect size perspective. That is, while 71–72% out of 100 pairs indicates a large difference between the groups, so does 28–29%. Third, the table can also be used in the opposite way: if we know that $d = 0.8$ and we select 100 pairs of cases from different ordinal subpopulations, we expect to see 72 such pairs where the higher-ranked case in g would score higher on the metric variable *X*. This would be labeled a "large" effect size. Fourth, although Metsämuuronen (2024b) provided a mechanism of generalized Cohen's *d* for the polytomous settings, the relationship between *PHD* and *d* is not as clear as it is for binary and dichotomous settings (see footnotes 7 and 8).

Conclusion and discussion

Main results in nutshell

The starting point for the article was the notion that the existing nonparametric effect size estimators such as Grissom-Kim *PS*, Cliff's *d*, and in some extent also Vargha-Delaney *A* are, in practical settings, limited to binary or dichotomous cases because of their dependence on *U*-test statistics. Instead, Somers' *D* can be used as the basis for effect sizes in both dichotomous and polytomous ordinal settings. The derivations primary led to two new, general estimators of ES called the "probability of higher subgroup dominance" based on either Somers' *D* ($PHD = 0.5 \times D + 0.5$) or Goodman-Kruskal *G* ($PHG = 0.5 \times G + 0.5$). These could be called "probability of higher subgroup superiority", "higher subgroup dominance statistic", or an index to "stochastic superiority of the higher subgroup" to mimic the epithets given to *PS*, and *A*. In addition, the derivations led to three modifications of the traditional estimators called the "*ordinal PS*", "*ordinal d*", and "*ordinal A*".

PHD and *PHG* are general estimators in two senses: first, unlike the traditional *PS* and *A*, they can be used naturally with polytomous settings and, second, *PS* and *A* (in their basic form) can be expressed by using *PHD*: $PS = A = 1 - PHD$. Thus, the formula for *PHD* allows the natural extension of $PS = A$ to polytomous ordinal settings. The estimates of *PHD* are somewhat more conservative than those of *PHG* because the latter omits the pairs with equal values in the metric variables in the calculation process.

It is worth noting that *PHD* and *PHG* can be used as the basis for a common language interpretation of *r* effect size, Cohen's *d*, and Cohen's *f* (see Metsämuuronen, 2024a), as well as for certain reliability estimates (Metsämuuronen, 2023c). In particular, Metsämuuronen (2024a) notes that when $PHD = 0.72$, i.e., when the effect size is "large", $d = 0.80$, $r = 0.37$, and $f = 0.40$ can all be interpreted in the same way: if we select 100 pairs of cases from different ordinal subpopulations, we would expect to see 72 such pairs in which the higher-ranked case in *g* would score higher on the metric variable *X*. Note that Metsämuuronen (2024a) does *not* discuss the use of *PHG* in the common language transformation of *r*, *d*, and *f*. However, the interpretation would be the same except for the note that "*if we consider only those pairs where we know which of the cases scored higher in X*".

General advantages of D and G as bases for ES

McGrath and Meyer (2006), Ruscio (2008), and Metsämuuronen (2022d), among others, have pointed out some of the challenges of correlation coefficients and the advantages of nonparametric estimators of association and effect size. McGrath and Meyer (2006) point out the challenges of the product-moment correlation coefficient (*r*), particularly the point-biserial correlation (r_{pb}), which is vulnerable when the number of cases in the categories of the dichotomous variable are not equal. This leads imprecision in Cohen *d*. Metsämuuronen (2024a, 2024b) suggests that sometimes the challenge lies in the simplified formulae used to transform *r* into *d*. The main challenge, however, is that *r* itself is prone to radical deflation, i.e., it gives too low values with respect to the true correlation, when the scales differ from each other (Metsämuuronen, 2022d).

Ruscio (2008) points to such advantages for *A* such as insensitivity to the base rate, i.e., unequal numbers of cases in the categories of the dichotomous variable, relevance to understanding treatment effects, ease of interpretation, and robustness to outliers and violations of parametric assumptions. These are also all advances of *PHD* and *PHG*, and they apply to all nonparametric estimators.

Based on simulations, Metsämuuronen (2022d) points out seven conditions, such as scale mismatch, number of categories in the variables, unequal proportions of cases in the categories, number of tied cases, and deviation from uniform distribution, that cause radical deflation of *r* and the coefficient eta (η), whereas *D* and *G* (and R_{PC}) are naturally deflation-free or close to it. Consequently, ES estimators based on *D* and *G* are more robust than those based on *r* or η. Metsämuuronen (2020a, 2021b) has summarized several advantages of *D* and *G* over more widely used correlation estimators such as *r*, η, and R_{PC}. Here, the advantages of *PHD* and *PHG* are discussed from this perspective. The traditional common language estimators *PS*, *d* and *A* are also discussed where appropriate. The discussion is summarized in Table 5.

First, *D* and *G* exactly reach the values +1 and −1, while *r*, η, and *RPC* cannot reach the limits when the scales of two variables differ from each other (*r* and η) or when traditional estimation procedures are used (R_{PC}). In particular, *r* and η are very sensitive to the discrepancy in the number of cases in the subpopulations (see algebraic reasons in Metsämuuronen, 2022b, 2022c and simulations in 2021b, 2022d; see also McGrath & Meyer, 2006; Ruscio, 2008): if the discrepancy in the number of cases is extreme, both *r* and η approach zero, regardless of the true difference between the means in the subpopulations. This leads to a radical underestimation of ES by Cohen's *d* and *f*, since these can be expressed by *r* and η (see discussion in Metsämuuronen,

Table 5. Comparison of traditional and new effect size estimators.

	estimators of association[1]			new CLES estimators[2]	traditional CLES estimators[3]		Cohen's ES estimators		explaining power[4]	
	D & G	*r & η*	R_{PC}	*PHD & PHG, ordinal PS, d, and A*	*PS & d*	*A*	*d*	*f*	R^2	η^2
1. accurately reaches the extreme values	x			x	x	x				
2. robust for non-linearity	x	[5]	x	x	x	x		x		x
3. natural correction for deflation	x		x	x	x	x				
4. directional nature	x	x	[6]	x	x	[7]	x	x	x	x
5. applicable to deterministic data sets	x			x	x	x				
6. concerns observed variables	x	x		x	x	x	x	x	x	x
7. conceptualizes the effect across the distribution and not just in the middle	x			x	x	x				
8. computationally simple	x	x		x	x	x	x	x	x	x
9. can be used in polytomous settings	x	x	x	x		[8]	x	x	x	x
10. can easily be used in meta-analysis because of the convertibility	x	x		x	[9]	[9]	x	x	x	x

[1] Somers' delta (*D*), Goodman–Kruskal gamma (*G*), Product-moment correlation coefficient (*R*), Coefficient eta (*η*), polychoric correlation (R_{PC}).
[2] Probability of the higher subgroup dominance based on Somers' delta (*PHD*), Probability of the higher subgroup dominance based on Goodman-Kruskal gamma (*PHG*).
[3] Grissom-Kim "probability of superiority" (*PS*), Cliff's "dominance statistic" (*d*), Vargha-Delaney "stochastic superiority" (*A*).
[4] Unadjusted R squared (R^2), eta squared (η^2).
[5] Of *r* and *η*, the latter is robust for linearity.
[6] The directional nature of R_{PC} is unknown. Metsämuuronen (2022d) discusses the matter and notes that it has directional properties although the algebraic proof is not obvious. Its algorithm refers to the directionality in the same direction as *R*.
[7] In the dichotomous case, the traditional *A* is a directional measure the same way as *PHD* and *D*. In the polytomous case, of the traditional variants of *A*, only A_{ord} is a directional measure.
[8] Formulae for polytomous and repeated settings are available for Vargha-Delaney *A,* but the estimates are not comparable to *PHD*, *PHG, and ordinal A*.
[9] The traditional estimators are easier usable in the meta-analytical process because they can be transformed into each other's scales. The polytomous variants of *A* are not convertible to other scales.

2022b, 2022c, 2023a, 2023b). The estimators of ES based on *D* or *G* (*PHD* and *PHG* as well as ordinal *PS, d,* and *A*) do not have the same kind of shortcomings related to apparent attenuation and deflation as estimators based on *r* and η, such as Cohen's *d* and *f* as well as R^2 and η^2.

Second, *D* and *G* are more robust to extreme observations and nonlinearity than *r* (Newson, 2002). As nonparametric coefficients based on ranks, *D* and *G* have the same strength in this respect as R_{PC}: the estimates are robust to the discrepancy in group sizes, scale widths, or distribution of the score variable (see simulations in Metsämuuronen, 2021b, 2022d). However, *G* and R_{PC} are more robust than *D* because the estimates of *D* are affected by the number of tied pairs, whereas *G* and R_{PC} are not (Metsämuuronen, 2021b, 2022d). Thus, the estimators of ES based on *D* or *G* (*PHD* and *PHG,* as well as ordinal *PS, d,* and *A*) are more robust than estimators based on *r* and η, such as Cohen's *d* and *f*, and R^2 and η^2. Of the *PHD* and *PHG*, the latter is robust to the number of tied cases.

Third, because the estimates of *D* and *G* (as well as R_{PC}) tend to be closer to the true association than those by *r* and η in dichotomous and binary settings and in polytomous settings up to 3 (*D*) or 4 (*G*) categories (see simulations in, e.g., Metsämuuronen, 2021a, 2021b, 2022d), *D* and *G* are superior to *r* and η as estimators of association in these settings. Therefore, in the binary and dichotomous settings and polytomous settings with less than 4 subpopulations in *g*, the estimators for ES based on *D* or *G* (*PHD* and *PHG* as well as ordinal *PS, d,* and *A*) tend to give estimates closer to the true value than the estimators based on *r* and η, such as Cohen's *d* and *f* as well as R^2 and η^2.

Fourth, *D* and *G* have a directional nature in the same direction as has η^2 in the general linear modeling settings (see Metsämuuronen, 2021a). In particular, *r* is also a directional coefficient in the same direction as η^2 (see Metsämuuronen, 2022b, 2022c; see also the discussion of the correct direction in *D* in Metsämuuronen, 2020a, 2021a, 2021b). In this respect, *G* is less ambiguous than *D*, because *G* reaches the meaningful direction of association *strictly* while *D* gives three options to choose from. In many cases, the directionality is an advantage in comparison over the truly symmetric estimators such as Kendall *tau-b*. Note that the directionality of R_{PC} and the traditional *A* is unknown. Metsämuuronen (2022d) discusses the matter with R_{PC} and notes that it has directional properties although the algebraic proof is not obvious. As for the traditional *A*, the directionality is less likely because the categories are combined in the process. The nondirectional property in the traditional *A* may be the reason why the estimate seems to be less extreme compared to ordinal *A*. Because of the directionality in the same direction, the interpretation of the estimators for ES based on *D* or *G* (*PHD* and *PHG* as well as ordinal *PS, d,* and *A*) is comparable to the other directional estimators based on *r* and η, such as Cohen's *d* and *f* as well as R^2 and η^2.

Fifth, unlike *r* and η, *D* and *G* can detect deterministically discriminating data sets. When $D = G = \pm 1$, all the cases in all subpopulations are in a deterministic ascending (or descending) order according to the order of the metric variable, regardless of the number of cases, the number of categories in the variables, the number of tied pairs, or the proportions of cases in different subpopulations obtained in the categorical variable. All these conditions strongly affect *r* and η, leading to deflation in the estimates of correlation (see simulations in Metsämuuronen, 2021b, 2022d) and, consequently, in the estimates of ES. In particular, *r* and η can reach the value ± 1 only if the number of categories in both variables is identical (see algebraic reasons in, e.g., Metsämuuronen, 2022b, 2022c). Therefore, estimators for ES based on *D* or *G* (*PHD* and *PHG* as well as ordinal *PS, d,* and *A*) provide robust estimates even with extreme data sets of a deterministic nature, in contrast to estimators based on *r* and η such as Cohen's *d* and *f* as well as R^2 and η^2.

Sixth, *D* and *G* refer to the observed variables in the analysis, which are easy to use in further research and to interpret in common language, whereas R_{PC} or other estimators based on inferred associations such as biserial or polyserial correlations refer to unknown, inaccessible, and

hypothetical variables, which are difficult to use in research (see the critical discussion on R_{PC} in Chalmers, 2018). Therefore, estimators of ES based on *D* or *G* (*PHD* and *PHG* as well as ordinal *PS, d,* and *A*) are particularly appropriate for the family of common language estimators of ES.

Seventh, *D* and *G* are applicable and robust with large, small, non-normal, and sparse data sets and cross-tables, whereas the applicability and accuracy of the estimation result of the *r*, *η*, and partly also R_{PC} depends on the form of the cross-tabulation and the normality of the phenomenon. This is related to Grissom's and Kim's (2001) observation that the nonparametric alternatives reconceptualize ES as an effect throughout a distribution, not just at its center. Thus, estimators of ES based on *D* or *G* (*PHD* and *PHG* as well as ordinal *PS, d,* and *A*) are robust even with small sample sizes and extremely non-normal data sets and throughout a distribution unlike the estimators based on *r* and *η* such as Cohen's *d* and *f* as well as R^2 and η^2.

Eighth, *D* and *G* are easy to calculate even manually in practical settings (see Excel template at https://doi.org/10.13140/RG.2.2.10199.18089), while the calculation of R_{PC}, for example, requires specific software packages and complex procedures. Therefore, the estimators for ES based on *D* or *G* (*PHD* and *PHG* as well as ordinal *PS, d,* and *A*) are easy to calculate and communicate with the users who are not familiar with complicated software packages or syntax. Once *D* or *G* has been calculated, *PHD* and *PHG* can be easily calculated by using common spreadsheet software. The estimates for *PHD* and *PHG* as well as for the ordinal *PS, d,* and *A,* can also be calculated by using an R package CLES (Niemensivu & Metsämuuronen, 2024).

Finally, a specific advantage of *PHD* and *PHG* (as well as ordinal *PS, d,* and *A*) over the traditional Grissom-Kim *PS,* Cliff's *d*, and different variants of *A* is that they can be used with polytomous ordinal settings without further modification. Of these, *PHD* and *PHG* treat the favorable cases in the same way as is usual in the settings associated with nonparametric correlations, i.e., that the higher category is more favorable than the lower value. In this respect, the ordinal *PS*, ordinal *d* and ordinal *A* treat the lower categories as more favorable. Moreover, when *D* or *G* are used in the analysis, they are usually reported. Therefore, we can easily use the proposed *PHD* and *PHG* (as well as ordinal *PS,* ordinal *d,* and ordinal *A*) in meta-analyses because they are easy to calculate without providing raw data in which to compute the estimates of *PS*, *A*, and Cliff's *d*. Peng and Chen (2014) note this as a specific shortcoming of the traditional *PS*, *A*, and Cliff's *d*. From a meta-analysis perspective, it is an advantage that regardless of which estimator—*D*, *PHD*, *ordinal PS*, *ordinal d*, *ordinal A*, and to some extent *G* and *PHG*—is known, the other estimators can be transformed into the same metrics in the same way that *r*, *d*, and *f* can be transformed into each other's scales.

Further elaboration of different variants of Vargha-Delaney a

Since the *ordinal A* is part of the family of different variants of *A*, and they have different usability, their differences and similarities are collected in Table 6. Among the variants, the stochastic homogeneity, A_{AAD}, and the pairwise stochastic equality, A_{AAPD}, are the most general or nonspecific statistics. Their multiple comparison is somewhat parallel to the omnibus *F*-test with η^2 or Cohen's *f* as the effect size, since all possible comparisons are made although a slightly different way in the estimators. Of these, the A_{AAPD} is more thorough than the A_{AAD} because it compares two groups at a time whereas the A_{AAD} combines the remaining groups. For example, with four subpopulations, the A_{AAD} is the mean of 4 comparisons whereas the A_{AAPD} is the mean of $4 \times 3/2 = 6$ comparisons.

A_{ik} is a specific measure for only one group of interest. A parallel parametric setting is obtained in the regression point displacement design and the associated regression analysis or two-group *F* test or *t* test with Cohen's *f* or *d*. Alternatively, A_{ik} could be used as an alternative to Cohen's f^2 which is used to assess the effect of single fixed effects.

A_{ord} and *ordinal A* are estimators for a directional ordinal setting, although the mechanism in the calculation of A_{ord}, which is based on the mean of successive *A* statistics which are based on the groupwise Mann-Whitney *U* statistics, makes it less specific than *ordinal A*, which is based on

Table 6. Variants of coefficient A in multiple populations settings.

variant	appropriate setting	principle of computing multiple comparison	specificity of the multiple comparison; parallel parametric setting with a post-hoc test	common language interpretation
traditional A	dichotomous settings	no multiple comparisons; based on Somers' D	not relevant	yes
stochastic homogeneity, A_{AAD}	polytomous nominal setting; cannot utilize the ordinal characteristics of the dataset	statistic *A* is computed for all sets of individual groups and the union of the other groups after which the mean of all *A* values forms the A statistic	general and nonspecific or conservative approach; indicates whether any of the groups differ from others; parallel to two-way F test with η^2 or Cohen's *f* as the effect size	no
pairwise stochastic equality, A_{AAPD}	polytomous nominal setting; cannot utilize the ordinal characteristics of the dataset	statistic *A* is computed for all pairs of groups, and these values are aggregated into a single estimate	general and nonspecific or conservative approach; indicates whether any of the groups differ from others; parallel to two-way F test with η^2 or Cohen's *f* as the effect size	no
specific subpopulation of interest, A_{ik}	polytomous nominal or ordinal settings	Groups of interest (*i*) are compared to the union of the other groups (*k*) *without* combining them as an aggregate statistic.	specific approach; indicates whether a single group differs from the others; parallel to two-group F test or t test with Cohen's f or d or Cohen's f^2 as the effect size	no
A_{ord}	polytomous ordinal settings	each ordinal subpopulation is compared with the closest (higher order) subpopulation by using *A*, after which the mean of these values is the estimate of *A*	less specific directional comparison of ordinal dataset; parallel to linear mixed models (LMM) and related marginal R^2 and conditional R^2 as the effect size	no
ordinal A	dichotomous or polytomous ordinal settings	based on directional Jonckheere-Terpstra test statistic and Somers' D	specific directional and conservative comparison of ordinal dataset; parallel to linear mixed models (LMM) and related marginal R^2 and conditional R^2 as the effect size	yes

directional Jonckheere-Terpstra test statistics and strict probability. For these types of directional settings, parametric parallels are linear mixed models (LMM) and related marginal R^2 and conditional R^2. Of the five alternatives for multiple comparison settings, *ordinal A* is the most specific; it is strictly directional, although somewhat conservative in that it is affected by the number of tied pairs as it is based on Somers' *D*. Of the variants, only the *ordinal A* and traditional *A* in dichotomous settings share a common language interpretation in the sense of McGraw and Wong (1992).

Restrictions and suggestions for further studies

Although *D* and *G* are traditional and easy-to-use association estimators for ordinal-ordinal variables or ordinal-metric variables, six limitations of the ordinal CLESs and related ASEs are discussed here.

First, if the metric variable is (semi-)continuous with large differences between the actual values, then converting it to ordinal categories obviously loses some information. This is typical of all the nonparametric estimators, not just *PHD* and *PHG* (as well as ordinal *PS,* ordinal *d,* and ordinal *A*).

Second, although not obvious, when the number of a categories in the categorical ordinal variable exceeds 3 (for *D*) or 4 (for *G*), the correlation between the ordinal *g* and the metric *X* estimated by using *D* or *G* tends to be lower than that obtained by *r* or η (see Göktaş & İşçi, 2011; Metsämuuronen, 2021b, 2022d). *D* and *G* do not underestimate the probability per se, but when the scales of two variables are large, the covariation (indicated by *r*) tends to be higher than the probability (indicated by *D* or *G*). This phenomenon is discussed by Kendall (1949), Newson (2002), and Metsämuuronen (2021a, 2021b, 2022d), and it can be explained by Greiner's relation (Greiner, 1909). With continuous variables *X* and *Y* (implying no tied pairs), $G = D$. In this case, Greiner's relation states that $\rho_{XY} = \sin\left(\frac{\pi}{2} \times D_{XY}\right) = \sin\left(\frac{\pi}{2} \times G_{XY}\right)$ (see Metsämuuronen, 2022d). Thus, with two continuous variables, if $G = D = 0.5$, $r = 1/\sqrt{2} = 0.7071$. Then, assuming continuous variables, except for the values ±1 and 0, the magnitudes of the estimates of *D* and *G* tend to be lower than those of *r*.

Third, related to the previous point, the more categories there are in the categorical ordinal variable, the more likely it is that the probability (estimated by *D* or *G*) will be lower than the correlation (estimated by *r*). This leads to the practical point that when the number of categories in the categorical ordinal variable exceeds 3 or 4, the magnitude of the effect size estimates from the estimators based on *D* or *G* are expected to be smaller in magnitude than when using the estimators based on *r* and η, such as Cohen's *d* and *f,* and R^2 and η^2. For larger scales, Metsämuuronen (2022a) suggests using the dimension-corrected *D* and *G* (D_2, G_2; Metsämuuronen, 2021b) instead of *D* and *G*. The potential of the estimators PHD_2 and PHG_2 would be worth further investigation.

Fourth, *PHD* and *PHG* (as well as ordinal *PS,* ordinal *d,* and ordinal *A*) are not the best options, nor are they suggested for use with a truly categorical polytomous nominal variable and a metric variable without first dummying the nominal variable as is done in the traditional *A*. Without dummying the nominal-scaled polytomous variable, Cohen's *f*, f^2, or η^2 are superior to *PHD* and *PHG,* and the traditional *A* is a relevant alternative for the parametric measures. Metsämuuronen (2024c) remind us, however, that in many cases the nominal-looking variable may have obvious or latent ordinal properties, and then the use of estimators based on *D* or *G* would be appropriate. Although we have estimators of ES for truly categorical variables such as the *phi* coefficient, Cramer's *V*, Cohen's *w*, and Cohen's *h*, these are not appropriate for a combination of a nominal-metric variables. For these settings, Freeman's theta (Freeman, 1965) may also be a coefficient to study further. As far as it is known, its properties as a common language estimator have not been studied yet. It is known, however, that when the categories are ordered, theta is a special case of Somers' delta (Hubert, 1974; Jacobson, 1972), although it was originally and primarily developed for use with nominal-ordinal or nominal-interval variables (see Freeman's reply to Hubert in Freeman, 1976). In fact, Jacobson (1972) showed that Freeman's theta is the absolute value of *D* although he was not aware of the direction. Metsämuuronen (2024c) shows that $\theta = |D(g|X)|$.

Fifth, since *D* and *G*, as being nonparametric estimators of association, are traditionally used with small samples, poorer scales, and non-normal populations, it is unsatisfactory that the estimation of standard errors is based on the normal approximation. It seems that in many cases the uniform approximation is a more natural option. We have formulae for testing null hypotheses assuming uniformity, but not for computing the ASE for confidence intervals. Even when using the uniform approximation for the null hypothesis, the statistical inference is based on a standardized normal distribution. The traditional bootstrap and jackknife techniques could be used to obtain the confidence intervals. However, these are not necessarily

applicable in meta-analysis settings where multiple statistics are combined from small samples. Simpler methods may be worth developing for these settings. Systematic studies in this regard would be beneficial.

Finally, the main limitation of the study is that there was no systematic simulation study of the new estimators *PHD* and *PHG*. The treatment was mainly conceptual and theoretical, and only a limited numerical example was given. Systematic studies of the behavior of *PHD* and *PHG* (as well as ordinal *PS,* ordinal *d,* and ordinal *A*) with respect to Cohen's *d* in polytomous settings would be beneficial.

If the existing R codes related to *PS*, *A* and *d* are restricted to binary or dichotomous settings only, it would be suggested to modify them those by also allowing their polytomous ordinal versions with established ASEs. For this purpose, the interested reader will find relevant R codes in the *CLES* package of Niemensivu and Metsämuuronen (2024).

Notes

1. From the point of view of *D* and *G*, the computational procedures do not care where the tied values are *within the variables g* and *X*, and the location has no effect on the result, only the ties within *pairs of observations* are of interest. In the direction "*g* given *X*", the discrepancy between the estimates of *D* and *G* is strictly dependent on the number of pairs where we do not know which of the cases in the categorical variable *g* (0 or 1 or 0 or 2) would score higher in *X*. The estimates of *D* and *G* will be identical when there are no such tied pairs. Otherwise, the estimate of *G* is larger than the estimate of *D*, and the discrepancy becomes larger the more such pairs are present in the data set (see Eqs. 16 and 17). The discussion of tied values and their location is more crucial for the estimates of the product-moment correlation coefficient (*r*): the number and location of tied cases *within a variable* have a strong effect on the deflation in the estimate. In particular, it is a well-known property of the point-biserial correlation (r_{pb}) that the difference in the proportions of 0s and 1s, i.e., the different number of tied cases in the categories of *g*, has a strong effect on the estimate (see, e.g., McGrath & Meyer, 2006; Metsämuuronen, 2022d; Ruscio, 2008). When the scales of two variables differ from each other, *r* never reaches the extremes of correlation (±1), the estimate is deflated, and the deflation approaches 100%, the more discrepancy there is in the number of (tied) cases in the variable with the narrower scale (see, e.g., Metsämuuronen, 2022d).
2. In fact, $T = 2T$, because the number of tied cases is calculated two times (see, e.g., Metsämuuronen, 2021a). Here, *T* is used as a general symbol for all tied pairs.
3. Ruscio's *RProbSup* package (2022), A1 function for two categories. Notably, at the time finalizing the article (01/30/2025), the estimate for $A_{g1} = 0.701$ although, logically, it should be $A_{g1} = 0.299$ as given by Ord1 function (see Footnote 5).
4. The first variant is called *stochastic homogeneity* (Vargha & Delaney, 1998) or the average absolute deviation (AAD), A_{AAD} in Ruscio and Gera (2013) and Ruscio (2022), associated with the Kruskal-Wallis test. In this variant, the *A* statistic is computed for all sets of individual subpopulations and the union of the other subpopulations, after which the mean of all *A* values forms the *A* statistic. The second variant is called the *pairwise stochastic equality* (Vargha & Delaney, 2000) or the average absolute pairwise deviation (AAPD), A_{AAPD} in Ruscio and Gera (2013) and Ruscio (2022). In this variant, the *A*-statistic is computed pairwise for all subpopulations, and these values are aggregated into a single estimate. When the number of subpopulations (*k*) exceeds 3, A_{AAPD} is based on more comparisons ($k(k–1)/2$) than A_{AAD} (*k*). Third, Ruscio and Gera (2013) discuss a variant in which a *specific subpopulation* is of interest, A_{ik} by Vargha and Delaney (1998) and reinterpreted in Ruscio and Gera (2013). In this method, the group of interest (*i*) is compared to the union of the other groups (*k*) *without* combining them as an aggregate statistic. From the viewpoint of the directional *D*, *PHD* and *PHG*, and *ordinal A*, the fourth variant of Ruscio and Gera (2013), A_{ord}, comes the closest. In this method, each ordinal subpopulation is compared with the closest (higher order) subpopulation, after which the mean of these values is the estimate of *A*. Using Ruscio's (2022) *RProbSup* package, with variable g2, $A_{AAD} = 0.759$, $A_{AAPD} = 0.837$, $A_{ik} = 0.487$, and $A_{ord} = 0.235$, whose interpretations are not obvious. None of the four variants of *A* leads to a "common language" interpretation in the sense of McGrath and Wong (1992), while $PHD = 0.832$, $PHG = 0.842$, and *ordinal* $A = 0.168$ do.
5. Ruscio's *RProbSup* package (2022), Ord1 function for two or more ordinal categories. Note that, at the time of finalizing the article (01/30/2025), the A1 and Ord1 functions give the probability in opposite ways. Using the A1 function, the estimate for the item g1 is $A_{g1} = 0.701$ and when using the Ord1 function, the estimate is $A_{g1} = 0.299$ and for g2, $A_{g2} = 0.235$, i.e., the A1 function gives (non-traditionally)

the probability that the *higher* categories are superior to the lower categories and the results with $A_{AAD} = 0.759$ and $A_{AAPD} = 0.837$ (see Footnote 4) suggest that the same is true for A_{AD} and A_{APD}. The Ord1 function gives the probability that the *lower* categories are superior to the higher categories.

6. The algorithm of Mehta and Patel (1983, 1986, 2013) was used in the IBM SPSS environment. It is worth noting that in the commonly used exact tests in the R environment, such as the *fisher.test* function of the *stats* package and the *fisher_test* function of the *rstatix* package using the Fisher-Freeman-Halton test (FFH; Fisher, 1925; Freeman & Halton, 1951), the crucial concepts are operationalized and computed in a different way than in the Mehta and Patel algorithm (MP; see Metsämuuronen, 2024c). In the cases of g1 and g2, both R functions give the value for the exact probability $p = 1.000$, which is obviously misleading considering the asymptotic p-values ($p = 0.067$ for g1 and $p < 0.001$ for g2). The latter probabilities make sense because the correlations are high. The reason for the discrepancy is that FFH uses a very ineffective discrepancy measure, the chi-squared statistic, while MP uses the more effective correlation coefficient. Thus, the result of FFH is not comparable to the asymptotic estimates and the MP for D and G. Metsämuuronen (2024c) points out that it is unnecessarily confusing that different algorithms give different exact probabilities for the same contingency table, and suggests changes in the algorithm of FFH.
7. Note that the thresholds for *PHD* and *PHG* depend somewhat on the discrepancy between the numbers of cases in the subpopulations. If the proportions of cases are equal, i.e. $p_i = p_j = 1/(\text{number of populations})$, the threshold is exact. If the absolute difference between the proportions of the smallest and largest subpopulations exceeds 0.40, the thresholds may be higher than those given in Table 4 (see Metsämuuronen, 2024a).
8. In fact, a better model between *PHD* and Cohen's d is based on Greiner's relation discussed by Metsämuuronen (2024a). The result is the same, except that the corresponding estimate for "medium" $d = 0.5$ would be $PHD = 0.64$, i.e., the same as given for Vargha-Delaney A (see Table 4). Metsämuuronen (2024b) also discusses a form of Cohen's d suitable for polytomous settings. This may make it possible to represent the thresholds also in the polytomous settings as well. However, in the polytomous settings, the thresholds depend on the number of subpopulations and the discrepancy between the proportions of the cases in the subpopulations (Metsämuuronen, 2024a). For the sake of simplicity, these are not discussed further here.

Acknowledgements

Sincere thanks to Professor, PhD, Salvatore Mangiafico at Rutgers, The State University of New Jersey for suggesting that we consider Freeman's theta and to Professor, PhD, Michael Smithson at Australian National University for reminding of the relationship between AUC and Somers D in a private discussion about the effect sizes. Also, sincere thanks to Counsellor of Evaluation Jukka Marjanen, Finnish Education Evaluation Centre, for providing preliminary Excel algorithms and PhD student Timi Niemensivu for R codes to calculate the estimates. The comments of two anonymous reviewers helped to tone the text and to find three new common language effect size estimators on the top of *PHD* and *PHG* that were originally proposed.

Ethic statement

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

Copyright statement

All necessary copyrights are in place for the research.

Preprint server

https://doi.org/10.13140/RG.2.2.14774.31045

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript. The study has been completed as part of the EDUCA Flagship project funded by the Research Council of Finland (#358924, #358947).

ORCID

Jari Metsämuuronen http://orcid.org/0000-0001-6027-0799

Data availability statement

The dataset used in the empirical section of the article is available in CSV format at https://doi.org/10.13140/RG.2.2.10530.76482 and in IBM SPSS format at https://doi.org/10.13140/RG.2.2.17594.72641.

References

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

AERA. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33–40. https://doi.org/10.3102/0013189X035006033

Berry, K. J., Johnston, J. E., & Mielke, Jr, P. W. (2018). *The measurement of correlation. A permutation statistical approach*. Springer. https://doi.org/10.1007/978-3-319-98926-6

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

Chalmers, R. P. (2018). On misconceptions and the limited usefulness of ordinal alpha. *Educational and Psychological Measurement*, *78*(6), 1056–1071. https://doi.org/10.1177/0013164417727036

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences* (1st ed.). Academic press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cureton, E. E. (1956). Rank–biserial correlation. *Psychometrika*, *21*(3), 287–290. https://doi.org/10.1007/BF02289138

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, *7*(4), 485–503. https://doi.org/10.1037/1082-989x.7.4.485

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Freeman, G. H., & Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, *38*(1–2), 141–149. https://doi.org/10.2307/2332323

Freeman, L. C. (1965). *Elementary applied statistics: For students in behavioral science*. John Wiley & Sons.

Freeman, L. C. (1976). A further note on Freeman's measure of association. *Psychometrika*, *41*(2), 273–275. https://doi.org/10.1007/BF02291845

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika*, *52*(1–2), 203–223. https://doi.org/10.1093/biomet/52.1-2.203

Gini, C. (1909). Concentration and dependency ratios (in Italian). *English Translation in Rivista di Politica Economica*, *87*(1997), 769–789.

Göktaş, A., & İşçi, O. A. (2011). Comparison of the most commonly used measures of correlation for doubly ordered square contingency tables via simulation. *Metodološki Zvezki [Methodological Notebooks]*, *8*(1), 17–37. https://www.stat-d.si/mz/mz8.1/goktas.pdf

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

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

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

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, *79*(2), 314–316. https://doi.org/10.1037/0021-9010.79.2.314

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, *6*(2), 135–146. https://doi.org/10.1037/1082-989x.6.2.135
Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). Routledge.
Harrell, F. (2001). *Regression modeling strategies*. Springer.
Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, *247*(18), 2543–2546. https://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
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
Hubert, L. (1974). A note on Freeman's measure of association for relating an ordered to an unordered factor. *Psychometrika*, *39*(4), 517–520. https://doi.org/10.1007/BF02291672
Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*(2), 227–240. https://doi.org/10.1177/0013164402062002002
IBM. (2024). *IBM SPSS statistics algorithms*. Retrieved January 24, 2025, from https://www.ibm.com/docs/SSLVMB_29.0.0/pdf/IBM_SPSS_Statistics_Algorithms.pdf
Jacobson, P. E. (1972). Applying measures of association to nominal-ordinal data. *The Pacific Sociological Review*, *15*(1), 41–60. https://doi.org/10.2307/1388286
Jansen, M. E. (1984). Ridit analysis, a review. *Statistica Neerlandica*, *38*(3), 141–158. https://doi.org/10.1111/j.1467-9574.1984.tb01106.x
Jonckheere, A. R. (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
Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. https://doi.org/10.1037/a0028086
Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1–2), 81–93. https://doi.org/10.2307/2332226
Kendall, M. G. (1948). *Rank correlation methods* (1st ed.). Charles Griffin & Co Ltd.
Kendall, M. (1949). Rank and product–moment correlation. *Biometrika*, *36*(Pt. 1-2), 177–193. https://doi.org/10.2307/2332540
Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). Oxford University Press.
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. https://doi.org/10.2307/2280779
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
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
McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods*, *11*(4), 386–401. https://doi.org/10.1037/1082-989x.11.4.386
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361
Mehta, C. R., & Patel, N. R. (1983). A network algorithm for performing Fisher's exact test in contingency tables. *Journal of the American Statistical Association*, *78*(382), 427–434. https://doi.org/10.1080/01621459.1983.10477989
Mehta, C. R., & Patel, N. R. (1986). ALGORITHM 643: FEXACT: A FORTRAN subroutine for Fisher's exact test on unordered r × c contingency tables. *ACM Transactions on Mathematical Software*, *12*(2), 154–161. https://doi.org/10.1145/6497.2143
Mehta, C. R., & Patel, N. R. (2013). *IBM SPSS Exact tests*. IBM Corp. Retrieved January 24, 2025, from https://www.ibm.com/docs/en/SSLVMB_27.0.0/pdf/en/IBM_SPSS_Exact_Tests.pdf
Metsämuuronen, J. (2020a). Somers *D* as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. *International Journal of Educational Methodology*, *6*(1), 207–221. https://doi.org/10.12973/ijem.6.1.207
Metsämuuronen, J. (2020b). 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. (2021a). Directional nature of Goodman-Kruskal gamma and some consequences. Identity of Goodman-Kruskal gamma and Somers delta, and their connection to Jonckheere-Terpstra test statistic. *Behaviormetrika*, *48*(2), 283–307. https://doi.org/10.1007/s41237-021-00138-8

Metsämuuronen, J. (2021b). 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
Metsämuuronen, J. (2022a). Rank–polyserial correlation: Quest for a "missing" coefficient of correlation. *Frontiers in Applied Mathematics and Statistics*, *8*, 914932. https://doi.org/10.3389/fams.2022.914932
Metsämuuronen, J. (2022b). Directional nature of the product–moment correlation coefficient and some consequences. *Frontiers in Psychology*, *13*, 988660. https://doi.org/10.3389/fpsyg.2022.988660
Metsämuuronen, J. (2022c). Artificial systematic attenuation in eta squared and some related consequences. Attenuation-corrected eta and eta squared, negative values of eta, and their relation to Pearson correlation. *Behaviormetrika*, *50*(1), 27–61. https://link.springer.com/content/pdf/10.1007/s41237-022-00162-2.pdf https://doi.org/10.1007/s41237-022-00162-2
Metsämuuronen, J. (2022d). Effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability. Seeking the best options of correlation for deflation-corrected reliability. *Behaviormetrika*, *49*(1), 91–130. https://doi.org/10.1007/s41237-022-00158-y
Metsämuuronen, J. (2022e). Typology of deflation-corrected estimators of reliability. *Frontiers in Psychology*, *13*, 891959. https://doi.org/10.3389/fpsyg.2022.891959
Metsämuuronen, J. (2022f). How to obtain the most error-free estimate of reliability? Eight sources of underestimation of reliability. *Practical Assessment, Research, and Evaluation, PARE*, *27*(1), 10. https://doi.org/10.7275/7nkb-j673
Metsämuuronen, J. (2023a). Note on the radical deflation in t-test statistic, some consequences, and deflation- corrected t-test statistic. Preprint. Retrieved January 24, 2025, from https://doi.org/10.13140/RG.2.2.25033.62564
Metsämuuronen, J. (2023b). Deflation-corrected F-test statistic, effect size, and explaining power. Note on the radical deflation in eta squared and F-test statistics. Preprint. Retrieved January 24, 2025, from https://doi.org/10.13140/RG.2.2.19260.51840
Metsämuuronen, J. (2023c). How to make sense to reliability? Common language estimators of reliability and the relation of reliability to effect size. Preprint. Retrieved January 24, 2025, from https://doi.org/10.13140/RG.2.2.31639.04000
Metsämuuronen, J. (2024a). Common language interpretation of r effect size, Cohen's d, and Cohen's f. Preprint. Retrieved January 24, 2025, from https://doi.org/10.13140/RG.2.2.14430.20804
Metsämuuronen, J. (2024b). Generalized Cohen d for polytomous settings. Preprint. Retrieved January 24, 2025, from https://doi.org/10.13140/RG.2.2.11970.96968/1
Metsämuuronen, J. (2024c). Note on the deviating exact probabilities. Should we change the logic of Fisher-Freeman-Halton test? Preprint. Retrieved January 24, 2025, from https://doi.org/10.13140/RG.2.2.34144.08967/1
Nelson, T. O. (1984). A comparison of current measures of feeling-of-knowing accuracy. *Psychological Bulletin*, *95*(1), 109–133. https://doi.org/10.1037/0033-2909.95.1.109
Newson, R. (1998). SOMERSD: Stata module to calculate Kendall's tau-a, Somers' D and median differences. *Statistical Software Components*, S336401, Boston College Department of Economics, revised 16 Apr 2020. https://ideas.repec.org/c/boc/bocode/s336401.html
Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal: Promoting Communications on Statistics and Stata*, *2*(1), 45–64. http://www.stata-journal.com/sjpdf.html?articlenum=st0007 https://doi.org/10.1177/1536867X0200200103
Newson, R. (2006). Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal: Promoting Communications on Statistics and Stata*, *6*(3), 309–334. http://www.stata-journal.com/sjpdf.html?articlenum=snp15_6 https://doi.org/10.1177/1536867X0600600302
Newson, R. (2008). Identity of Somers' *D* and the rank biserial correlation coefficient. Retrieved January 24, 2025, from http://www.rogernewsonresources.org.uk/miscdocs/ranksum1.pdf
Neyman, J., & Pearson, E. S. (1933a). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289–337. https://doi.org/10.1098/rsta.1933.0009
Neyman, J., & Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, *29*(4), 492–510. https://doi.org/10.1017/S030500410001152X
Niemensivu, T., & Metsämuuronen, J. (2024). CLES. R package for calculating common language effect sizes. Retrieved January 24, 2025, from https://github.com/TNiemensivu/CLEF
Ortelli, O. A. (2022). RProbSup: Calculating the probability of superiority. *TCNJ Journal of Student Scholarship*, *24*, 1–7. https://joss.tcnj.edu/wp-content/uploads/sites/176/2022/04/2022-Ortelli-Psychology.pdf
Peng, C.-Y. J., & Chen, L.-T. (2014). Beyond Cohen's d: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education*, *82*(1), 22–50. https://doi.org/10.1080/00220973.2012.745471
Ricca, B. P., & Blaine, B. E. (2022). Brief Research Report: Notes on a nonparametric estimate of effect size. *The Journal of Experimental Education*, *90*(1), 249–258. https://doi.org/10.1080/00220973.2020.1781752
Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Ruscio, J. (2022). RProbSup: Calculates Probability of Superiority. R package version 3.0. Retrieved January 24, 2025, from https://CRAN.R-project.org/package=RProbSup
Ruscio, J., & Gera, B. L. (2013). Generalizations and extensions of the probability of superiority effect size estimator. *Multivariate Behavioral Research*, *48*(2), 208–219. https://doi.org/10.1080/00273171.2012.738184
Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, *47*(2), 201–223. https://doi.org/10.1080/00273171.2012.658329
Sawilowsky, S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599. https://doi.org/10.22237/jmasm/1257035100
Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813. https://doi.org/10.3389/fpsyg.2019.00813
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 behavioral sciences* (2nd ed.). McGraw-Hill.
Sirkin, M. R. (2006). *Statistics of the social science* (3rd ed.). SAGE Publications.
Smithson, M. (2023). The receiver operating characteristic area under the curve (or Mean ridit) as an Effect Size. *Psychological Methods*. https://doi.org/10.1037/met0000601
Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, *27*(6), 799–811. https://doi.org/10.2307/2090408
Somers, R. H. (1980). Simple approximations to null sampling variances. Goodman and Kruskal's gamma, Kendall's tau and Somers dyx. *Sociological Methods & Research*, *9*(1), 115–126. https://doi.org/10.1177/004912418000900107
Stata Corp. (2018). *Stata manual*. Retrieved January 24, 2025 from https://www.stata.com/manuals13/r.pdf
Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae (Proceedings)*, *55*(3), 327–333. https://doi.org/10.1016/S1385-7258(52)50043-X
Thompson, B. (1988). A note on statistical significance testing. *Measurement and Evaluation in Counseling and Development*, *20*(4), 146–148. https://doi.org/10.1080/07481756.1988.12022864
Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development*, *22*(1), 2–6. https://doi.org/10.1080/07481756.1990.12022905
Thompson, B. (1993). Foreword. *The Journal of Experimental Education*, *61*(4), 285–286. https://doi.org/10.1080/00220973.1993.10806590
Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837–847.
Torchiano, M. (2020). R package *effsize*. Retrieved January 24, 2025 from https://cran.r-project.org/web/packages/effsize/index.html
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
Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, *23*(2), 170–192. https://doi.org/10.3102/10769986023002170
Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. https://doi.org/10.3102/10769986025002101
Wendt, H. W. (1972). Dealing with a common problem in social science: A simplified rank biserial coefficient of correlation based on the U statistic. *European Journal of Social Psychology*, *2*(4), 463–465. https://doi.org/10.1002/ejsp.2420020412
Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (Eds.) (2015). *Handbook of practical program evaluation* (4th ed.). Jossey-Bass.
Wilkinson, L., and Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594
Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968
Woods, C. M. (2007). Confidence intervals for gamma-family measures of ordinal association. *Psychological Methods*, *12*(2), 185–204. https://doi.org/10.1037/1082-989X.12.2.185

Appendix A. Grissom–Kim PS, Cliff's d, and Vargha–Delaney A as special cases of Somers D

Grissom–Kim PS as a transformation of Somers D

Grissom–Kim *PS* or "probability of superiority" (Grissom, 1994; Grissom & Kim, 2001) can be expressed as follows:

$$PS = \frac{U_0}{n_0 n_1} \tag{A1}$$

(e.g., Peng & Chen, 2014), where U_0 is the Wilcoxon–Mann–Whitney statistic related to the lower of the subpopulations 0 and 1 as discussed above. That *U* in Equation (1 in the main text) is specifically related to the subpopulation 0 is not specifically defined in the literature, but the result of the estimation shows this. This is discussed below. Note that in this form, *PS* is restricted to dichotomous cases because the embedded *U* statistic is used with dichotomous settings. Later, a simple modification of *PS* is derived that can be also used with polytomous ordinal cases.

Comparing Equations (11 in the main text) and (A1), we know that *PS* is strictly related to Somers' delta (*D*; Somers, 1962) as follows:

$$1 - 2PS = 1 - 2 \times \frac{U_0}{n_0 n_1} = D\big(g|X\big) \tag{A2}$$

In respect to *PS*, *D* is directed so that the variable with the larger scale (*X*) explains the order in the variable with the shorter scale (*g*), i.e., $D(g|X)$ (see the arguments for the proper direction in Metsämuuronen, 2020). This direction is traditionally worded as "*X* dependent" in the settings related to general linear modeling or "*g* given *X*" in the settings related to conditional probabilities and measurement modeling. Consequently,

$$PS = 1 - \big(0.5 \times D\big(g|X\big) + 0.5\big). \tag{A3}$$

Hence, Grissom–Kim *PS* in the form of Equation (1) is a modification and a special case of Somers' *D* directed so that "*X* dependent" or "*g* given *X*" in dichotomous settings.

Cliff's *d* as a transformation of Somers' Delta

Cliff's *d* or "dominance statistic" (Cliff, 1993) can be expressed as

$$Cliff's\ d = \frac{(X_0 > X_1) - (X_0 < X_1)}{n_0 n_1}, \tag{A4}$$

where # is the number of times that the argument in the parentheses is true in the data set, and $n1$ and n_2 are the sample sizes in groups 0 and 1. In particular, from the perspective of the nonparametric association estimators discussed above, $(X_0 > X_1)$ refers to the number of discordant pairs (*Q*) and $(X_0 < X_1)$ refers to the number of the concordant pairs (*P*). An alternative form of *d* is

$$Cliff's\ d\ = \frac{2U_0}{n_0 n_1} - 1, \tag{A5}$$

(e.g., Peng & Chen, 2014) where U_0 is the Wilcoxon–Mann–Whitney statistic associated with the *lower* of the groups of 0 and 1. As with Grissom–Kim *PS*, the nature of the *U* statistic is not strictly defined in the literature, but the value is obvious from the result.

In the form of Equation (A5), Cliff's *d* is restricted to binary cases because of its interdependence with the *U* statistic. Notably, Equation (A5) *looks* the same as Cureton's rank-biserial correlation based on the proportion of favorable cases (Equation A3 in the main text), but where *U* refers to the *higher* of the groups of 0 and 1. Thus, Equation (A5) leads us to an opposite value

than Cureton's formula (Equation 3 in the main test) as well as Wendt's (Equation 4 in the main text), the latter also based on the lower of the groups of 0 and 1. Hence, *Cliff's* $d = -R_{RB}$: if $R_{RB} = +1$, Cliff's $d = -1$ and vice versa.

Newson (2008) showed that Cureton's formula is a special case of Somers' delta (D). Hence, in the form of Equation (A5), Cliff's d is a transformation and a special case of D: a reverse of Somers' D directed so that "X dependent" restricted to binary or dichotomous settings:

$$\text{Cliff's } d = -\frac{P - Q}{n_1 n_2} = -D(g|X) \tag{A6}$$

(see Equation 10 in the main text). Notably, because of Equation (A3), Cliff's d can also be expressed as follows:

$$\text{Cliff's } d = 1 - 2 \times \left(0.5 \times D(g|X) + 0.5\right). \tag{A7}$$

As with Grissom–Kim *PS*, Cliff's d is restricted to the dichotomous cases, whereas Somers' D can be used with both dichotomous and polytomous ordinal cases. Thus, D could be used as an indicator of "dominance statistic" instead of Cliff's d. This option is discussed later. From the point of view of interpreting the effect size, it would be suggested to use the logic of the nonparametric correlation indices: the positive value refers to the cases where the group with the higher category value would be labeled as the "favorable case" and not the opposite. As noted above, the estimate of Somers' D can be easily transformed into a form that indicates the proportion of the logically (ascending) located observations in g after ordering them by the score X (Equation A3).

Vargha–Delaney A as a transformation of Somers' Delta

In contrast to *PS* and d, Vargha and Delaney (2000) have provided us with several variations of A for different settings including a polytomous one. In the dichotomous case, Vargha–Delaney A or "stochastic superiority" can be expressed as

$$\textit{Vargha–Delaney } A = \frac{(X_0 > X_1) + 0.5(X_0 = X_1)}{n_0 n_1}, \tag{A8}$$

where # is the number of times that the argument in the parentheses is true in the data set, and n_1 and n_2 are the sample sizes in groups 0 and 1. An alternative form is

$$\textit{Vargha–Delaney } A = \frac{\frac{R_0}{n_0} - \left(\frac{n_0+1}{2}\right)}{n_1}, \tag{A9}$$

(e.g., Peng & Chen, 2014) where R_0 is the rank sum of the scores associated with the subpopulation 0, ranked among all (n_1+n_2) scores, that is, the Wilcoxon statistic W_0 (Wilcoxon, 1945) related to the lower of the subpopulations 0 and 1. Notably, the computational form of Mann–Whitney–Wilcoxon U statistic uses $W_0 = R_0$ as the basis for calculating the statistic, i.e., $U_0 = R_0 - (\frac{n_0(n_0+1)}{2})$ and $U_1 = R_1 - (\frac{n_1(n_1+1)}{2})$ (see Siegel & Castellan, 1988). Equation (A9) can be expressed in a form $A = \frac{R_0 - (\frac{n_0(n_0+1)}{2})}{n_0 n_1}$, where the numerator is equal to the U statistic associated with the lower of the groups of 0 and 1, i.e., U_0. Therefore, based on Equations (A9) and (11 in the main text), Vargha–Delaney A can be expressed as

$$A = \frac{U_0}{n_0 n_1} = PS. \tag{A10}$$

Thus, Vargha–Delaney A is algebraically identical to Grissom–Kim *PS* and, in the form of Equation (A10), both are restricted to dichotomous settings when expressed in this form because of their interdependence on U statistics.

Because of Equation (11 in the main text), the rank-biserial correlation formula R_{RB} can be expressed by using Wendt's formula (1972) by using the Wilcoxon test statistic, i.e., by the sum of ranks, as follows:

$$R_{RB} = 1 - 2 \times \frac{U_0}{n_0 n_1} = 1 - 2 \times \frac{R_0 - \left(\frac{n_0(n_0+1)}{2}\right)}{n_0 n_1} = 1 - 2 \times \frac{\frac{R_0}{n_0} - \frac{(n_0+1)}{2}}{n_1} = D\big(g|X\big) \tag{A11}$$

by using the lower of the two groups as the base. Because of Equations (A9) (A10), and (A11), in the dichotomous case, Vargha–Delaney A can be expressed as

$$\text{Vargha–Delaney } A = \frac{\frac{R_0}{n_0} - \left(\frac{n_0+1}{2}\right)}{n_1} = -0.5D\big(g|X\big) + 0.5 = 1 - \big(0.5 \times D\big(g|X\big) + 0.50\big) \tag{A12}$$

Thus, A in the form of Equations (A9) and (A10) is a modification and a special case of Somers' D the same way as PS: directed so that "X dependent" or "g given X" in dichotomous settings. As with PS and d, the estimates of A are reversed from the traditional notion of favorable cases: highly positive D becomes close to $A = 0$ and highly negative D becomes close to $A = +1$ the same manner as with Grissom–Kim PS.

Note that the statistic $0.5 \times D(g|X) + 0.50$ in Equations (A3) (A7), and (A12) is a new effect size estimator "probability for higher subgroup dominance based on Somers D", PHD, discussed in the main article. Thus, in the dichotomous settings, Vargha-Delaney A can be expressed by PHD: $A = 1 - PHD$. However, Vargha and Delaney's (2000) solution for the polytomous settings does not lead to relation with D and PHD. This is discussed in the main text.

References

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79*(2), 314–316. https://doi.org/10.1037/0021-9010.79.2.314

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6*(2), 135–146. https://doi.org/10.1037/1082-989x.6.2.135

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

Newson, R. (2008). Identity of Somers' D and the rank biserial correlation coefficient. Retrieved September 20, 2024, from http://www.rogernewsonresources.org.uk/miscdocs/ranksum1.pdf

Peng, C.-Y. J., & Chen, L.-T. (2014). Beyond Cohen's d: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education, 82*(1), 22–50. https://doi.org/10.1080/00220973.2012.745471

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

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

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*(2), 101–132. https://doi.org/10.3102/10769986025002101

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

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