Psychometrika (2026), **91**, 1069–1091
doi:10.1017/psy.2026.10085

PSYCHOMETRIC SOCIETY

LITERATURE REVIEW

Quadratically Weighted Agreement Coefficients: Interpretations and Connections

Rutger van Oest[1] and Jonas Moss[2]

[1]Department of Marketing, BI Norwegian Business School, Norway; [2]Department of Data Science and Analytics, BI Norwegian Business School, Norway

Corresponding author: Rutger van Oest; Email: rutger.d.oest@bi.no

(Received 26 January 2025; revised 3 November 2025; accepted 11 January 2026)

Abstract

In this study, we review interpretations and connections of chance-corrected agreement coefficients with quadratic weights, applicable when raters classify objects or subjects on an ordinal scale. Whereas correlation interpretations exist for coefficients with noninterchangeable raters, represented by Cohen's two-rater and Conger's multirater kappas, interpretations are essentially absent for interchangeable raters, represented by Fleiss' kappa and its two-rater version, Scott's pi. We show that Fleiss' quadratically weighted kappa equals Lin's generalized concordance correlation coefficient after recentering the rater means and covariance matrix at the grand mean. Furthermore, it equals the Pearson product–moment correlation and associated regression slope for two random raters, computed after concatenating the ratings from all possible rater pairs, including reversed rater order. Next, we demonstrate that Fleiss' and Conger's quadratically weighted kappas are linear transformations of each other, entirely determined by the pairwise differences in rater means and corresponding variances. As these kappas coincide if the rater means coincide, the conceptual distinction between interchangeable and noninterchangeable raters becomes empirically irrelevant if raters have (approximately) the same mean rating, even with substantially different rating distributions. Finally, Fleiss' quadratically weighted kappa (i.e., interchangeable raters) cannot exceed Conger's (i.e., noninterchangeable raters).

Keywords: concordance correlation; Conger's kappa; Fleiss' kappa; interrater agreement; Pearson product–moment correlation

A common and subjective task is classifying objects or subjects, hereafter called items, into predefined, mutually exclusive categories. Examples include content analysis that classifies subjective communications into content categories, meta-analysis that groups empirical studies based on underlying study settings, and medicine that requires appropriate diagnoses of patients. As subsequent analyses and conclusions critically depend on the obtained categorical data, validation is needed. It is, therefore, good practice to assign multiple independent raters and compute the percentage of cases in which these raters chose the same category and thus agreed, where a low percent agreement suggests discarding the data and improving the process. However, researchers also recognized early that this percent agreement is often substantially inflated: It counts not only genuine agreements but also chance agreements in which at least one rater was unsure and coincidently guessed the same category as some other rater, which may happen due to limited choice options. These notions have resulted in chance-corrected agreement coefficients. We provide a review with novel interpretations and connections for so-called quadratic weights.

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

1. Chance-corrected agreement coefficients

Cohen's kappa and Scott's pi are two classic agreement coefficients for two raters (Cohen, 1960; Scott, 1955). The chance correction in Cohen's kappa assumes that raters use their *own* category proportions (computed from the data) when guessing an item's category. In contrast, Scott's pi assumes that the guessing probabilities are the *same* for both raters by averaging the rater-specific category proportions over the two raters. When extending beyond two raters (i.e., moving to multirater coefficients), Cohen's kappa generalizes into Conger's kappa (Conger, 1980), and Scott's pi generalizes into Fleiss' kappa (Fleiss, 1971). Both multirater coefficients base the percentage of *observed* agreement on all items and different rater pairs. However, whereas Conger's kappa computes the probability of *chance* agreement by averaging this probability across all rater pairs using *rater-specific* category proportions, Fleiss' kappa uses the *same* (averaged) category proportions for all raters.

Conceptually, the main difference between Scott's pi and Fleiss' kappa, on the one hand, and Cohen's and Conger's kappas, on the other hand, is that the former two coefficients consider raters to be interchangeable (by averaging the raters' category proportions), whereas the latter two coefficients do not (Banerjee et al., 1999; Krippendorff, 2004). In addition to Scott's pi and Fleiss' kappa, coefficients with interchangeable raters include Krippendorff's alpha (Krippendorff, 2018) and the recently introduced uniform prior coefficient (Van Oest, 2019; Van Oest & Girard, 2022). The latter two coefficients involve traditional versus Bayesian small-sample corrections and converge to Fleiss' kappa. In the alternative class of noninterchangeable raters, Light's and Hubert's kappas complement those of Cohen and Conger (Hubert, 1977; Light, 1971). Light's version is the average Cohen's kappa across all rater pairs (Conger, 1980). In contrast, Hubert's version considers *simultaneous* agreement among all raters without breaking them up into pairs. It coincides with Conger's kappa if we measure the degree of *simultaneous* agreement by the fraction of rater pairs that result in agreement. More generally, the weighted counterparts of these two coefficients coincide if the weight for simultaneous agreement becomes the corresponding average weight across all rater pairs, which is a well-known choice (Mielke et al., 2007, 2009; Schuster & Smith, 2005; Warrens, 2012a).

Noninterchangeable raters are suitable when all raters classify all items and particularly natural when these raters are fixed (i.e., deliberately selected), such as specific experts, devices, or coding models that constitute the population of interest. In contrast, coefficients with interchangeable raters are needed when raters vary across items (Fleiss, 1971; Janson & Olsson, 2004; Landis & Koch, 1977a; Vanbelle, 2019). However, the preferred option is less clear-cut when all raters classify all items while being randomly selected to represent a larger population. Zwick (1988) proposed empirically investigating whether raters vary in category proportions beyond random noise to decide whether raters are interchangeable. A well-known "paradox" of Cohen's kappa, representing noninterchangeable raters, is that, *given the percentage of observed agreement*, this kappa increases as raters deviate more in their category proportions, suggesting that rater bias gets rewarded (Feinstein & Cicchetti, 1990). However, a counterargument is that the involved *ceteris paribus assumption* is unrealistic: A change in classification outcomes that triggers additional rater bias typically also decreases the percentage of observed agreement and ultimately Cohen's kappa, mitigating or even resolving the issue (Hoehler, 2000).

All so-far discussed agreement coefficients are *distribution-based*, where the chance correction uses raters' category proportions. An influential alternative class consists of *category-based* coefficients, where the chance correction uses only information on how many categories and assumes equal category probabilities when raters guess an item's category. The *S*-coefficient represents this second class (Bennett et al., 1954). Other researchers rediscovered this coefficient via alternative but ultimately equivalent approaches (Brennan & Prediger, 1981; Byrt et al., 1993; Holley & Guilford, 1964; Janson & Vegelius, 1979; Maxwell, 1977). A third class that has become important is *Gwet's framework*, yielding the AC1 and AC2 coefficients (Gwet, 2008, 2014). Figure 1 summarizes all coefficients to guide stakeholders, such as applied researchers, psychometricians, statisticians, and other professionals. The scope of the present study is the distribution-based class, including Cohen's and Fleiss' kappas, the most frequently used and cited two-rater and multirater coefficients (e.g., Andres & Hernandez, 2020; Vanbelle, 2019).

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

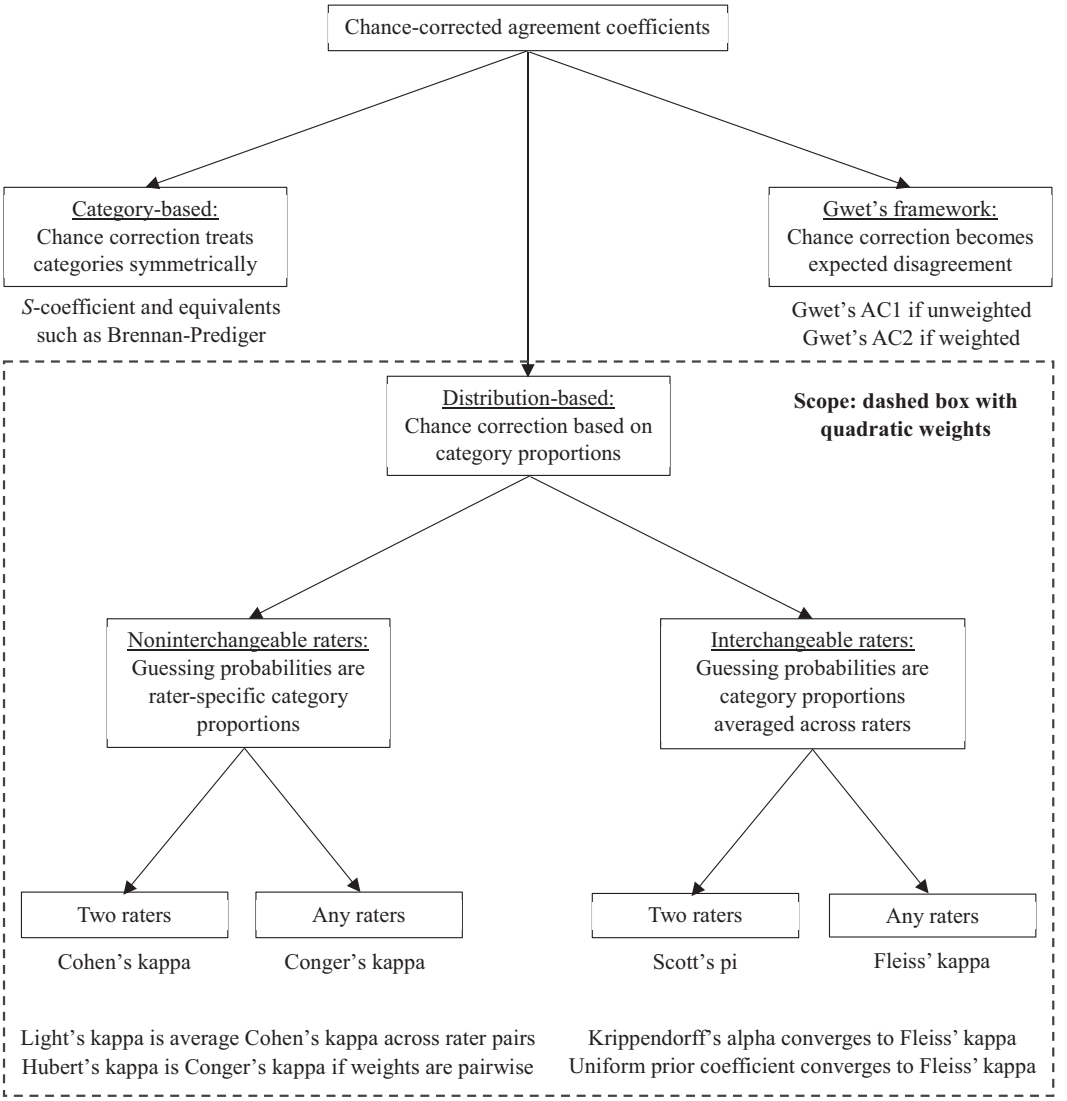


Figure 1. Overview of chance-corrected agreement coefficients plus the scope of the present study.

A second well-known "paradox" of Cohen's kappa (i.e., distribution-based coefficients) is that a high observed agreement among raters may still result in a low coefficient value if the categories have substantially unequal proportions (Feinstein & Cicchetti, 1990). However, others argued that Cohen's kappa does what it is supposed to do (Vach, 2005; Zwick, 1988). Typically, researchers want raters to agree not only on highly prevalent categories (e.g., not having a disease) but particularly on rare categories (e.g., having that disease). A situation in which raters usually agree on highly prevalent categories but never on rare categories may have serious consequences. However, the percent observed agreement, *S*-coefficient, and Gwet's coefficient would take high values. The *S*-coefficient linearly transforms the observed agreement using only the number of categories without other information (Zwick, 1988); Gwet's coefficient compares the observed agreement with the corresponding expected *disagreement*, resulting in higher values as the category proportions become more unequal (Vach & Gerke, 2023).

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

2. Incorporating weights for partial disagreements

All coefficients allow for generalization to ordinal (ordered) categories (Cohen, 1968; Gwet, 2014). For example, content analysis may require communication ratings on a five-point scale from very negative (1) to very positive (5). In such situations, some rater disagreements are worse than others, requiring a weighting scheme to describe the amount of penalization for each possible rater disagreement. Typically, exact agreement (i.e., both raters choose the same category) receives a disagreement weight of zero, and the maximum possible disagreement corresponds to a weight of one. All other disagreements receive either the maximum weight of one (reducing to the unweighted case of nominal categories) or a lower distance-based weight, reflecting *partial* disagreement. Two common weighting schemes are linear and quadratic (Cicchetti & Allison, 1971; Fleiss & Cohen, 1973). The former penalizes rater disagreements in proportion to the distance between the chosen categories; the latter penalizes based on the squared distance. A third, less common, weighting scheme uses radical weights, the distance's square root (Gwet, 2014). Thus, linear, quadratic, and radical weights penalize disagreements by taking the involved distance (expressed as a fraction of the maximum distance) to the power 1, 2, or 0.5. Quadratic weights are most lenient to partial disagreements and usually yield the highest coefficient values (Van Oest, 2023; Warrens, 2012b, 2013), potentially explaining their popularity in empirical studies (Vanbelle, 2016). However, quadratic weights also offer valuable interpretations as correlation coefficients. All three weighting schemes assume equidistant ordinal categories, where adjacent categories are always the same one point apart. Our results require this well-accepted assumption to hold.

Before turning to quadratic weights, we briefly review existing results for *linear* weights. First, Cohen's linearly weighted kappa is a weighted average of kappas computed for all possible reductions of the considered ordinal scale (e.g., five-point) into binary scales by merging adjacent categories (Vanbelle & Albert, 2009; Warrens, 2011). Second, Cohen's linearly weighted kappa is a weighted average of all linearly weighted kappas computed after merging any two adjacent categories in the ordinal scale, effectively reducing the number of categories by one (Warrens, 2012c). Third, these results generalize to any partition of the ordinal scale by merging adjacent categories: Cohen's linearly weighted kappa is a weighted average of linearly weighted kappas computed for all possible scale partitions given the chosen number of categories. Furthermore, invariance to the number of categories immediately extends this result to the weighted average across all non-trivial partitions (excluding only the original scale and its reduction to one single category); the weights always correspond to the denominator of the involved kappa elements (Warrens, 2012c). Fourth, Kvålseth (2018) showed that Cohen's linearly weighted kappa equals the corresponding unweighted kappa if cumulative probabilities replace the original probabilities when computing the latter. Fifth, Cohen's linearly weighted kappa captures the first moment of the distribution of rating distances generated by the two raters: One minus kappa equals the expected observed rating distance, using the bivariate distribution of ratings, expressed as a fraction of the rating distance that would be expected by chance, under independence (Vanbelle, 2016). These results for Cohen's kappa with linear weights yield elegant decompositions and interpretations, but they do not lead to correlation coefficients.

In contrast, Cohen's *quadratically* weighted kappa converges to two-factor intraclass correlation coefficients with random and fixed raters (i.e., the so-called ICC(2,1) and ICC(3,1) coefficients), where the items and raters are the two factors (Carrasco, 2010; Fleiss & Cohen, 1973; Janson & Olsson, 2001; Schuster, 2004; Schuster & Smith, 2005; Shrout & Fleiss, 1979; Warrens, 2014). Furthermore, this kappa coefficient equals Lin's concordance correlation coefficient (Andres & Hernandez, 2020; Carrasco, 2010; King & Chinchilli, 2001; Kvålseth, 2018; Schuster, 2004). Lin's concordance correlation possesses several attractive properties and measures agreement (not mere association) between two raters when the ratings are on a continuous scale (Barnhart et al., 2002; Lin, 1989). In addition, Cohen's quadratically weighted kappa reduces to the Pearson product–moment correlation if the two raters produced ordinal ratings with the same mean and variance (Brusco et al., 2008; Cohen, 1968; Schuster, 2004). For Conger's kappa, the multirater version of Cohen's kappa, quadratic weights imply convergence to two-factor intraclass correlation coefficients with random and fixed raters, and equivalence to Lin's *generalized*

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

concordance correlation coefficient that allows for any number of raters (Andres & Hernandez, 2020; Barnhart et al., 2002; Carrasco & Jover, 2003; Lin, 1989, 2000; Schuster & Smith, 2005).

Despite these helpful interpretations, quadratic weights have also received criticism, such as making Cohen's kappa sensitive to the number of included categories (Brenner & Kliebsch, 1996), making it behave like a coefficient of association instead of agreement (Graham & Jackson, 1993), and making it non-responsive to changes in the observed exact agreement on the middle category in situations with an odd number of categories combined with at least one rater having a symmetric distribution of category proportions (Warrens, 2012d).

Although interpretations are abundant for Cohen's and Conger's kappas, they are scarce for Fleiss' kappa, the most frequently used and cited *multirater* agreement coefficient. Gwet (2014, p. 67) noted that "there is no known procedure that formally generalizes Scott's coefficient to Fleiss'," and Stoyan et al. (2018, p. 381) referred to "Fleiss' κ, which is known to be difficult to interpret." Still, one existing result is that Fleiss' kappa converges to the one-factor intraclass correlation coefficient, where the items constitute the only factor due to interchangeable raters (Banerjee et al., 1999; Janson & Olsson, 2004; Landis & Koch, 1977a; Vanbelle, 2019). Furthermore, a model-based interpretation is that Fleiss' kappa estimates the probability that both raters in a pair assign an item to its correct category without guessing (Moss, 2023; Van Oest, 2019; Van Oest & Girard, 2022). However, interpretations in terms of concordance and Pearson product–moment correlations remain absent.

The present study addresses this literature gap and provides novel interpretations of Fleiss' quadratically weighted kappa (and its two-rater version, Scott's pi). We write it as Lin's generalized concordance correlation coefficient after recentering the raters' mean vector and covariance matrix at the grand mean to reflect interchangeable raters. Furthermore, Fleiss' quadratically weighted kappa equals the Pearson product–moment correlation and the regression slope coefficient after concatenating the ratings from all different rater pairs (including reversed rater order) into two data columns. Next, we show a linear connection between Fleiss' and Conger's quadratically weighted kappas, entirely determined by the raters' means and variances, and prove that Fleiss' quadratically weighted kappa never exceeds Conger's. The two coefficients coincide if the raters share the same mean rating without needing the same rating proportions or variance. All results extend to Scott's pi and Cohen's kappa for two raters with quadratic weights. Thus, we present new interpretations and connections for distribution-based coefficients (see the dashed box in Figure 1).

3. Two noninterchangeable raters: Cohen's quadratically weighted kappa

To establish ideas, we reproduce that Cohen's kappa, with quadratic weights, is equivalent to Lin's bivariate concordance correlation and reduces to the Pearson product–moment correlation if the two raters share the same mean and variance. Cohen's weighted kappa equals one minus the ratio of observed weighted disagreement, d_o, and expected-by-chance weighted disagreement, d_e (Cohen, 1968). Expressed in population parameters with C ordered categories, it is

$$\kappa = 1 - \frac{d_o}{d_e} = 1 - \frac{\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} v(c,\tilde{c}) \Pr(Y_1 = c, Y_2 = \tilde{c})}{\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} v(c,\tilde{c}) \Pr(Y_1 = c) \Pr(Y_2 = \tilde{c})}, \tag{1}$$

where $v(c,\tilde{c})$ denotes the disagreement weight when the first rater chooses category c and the second rater chooses category $\tilde{c}$. Furthermore, Y_1 and Y_2 denote the raters' category choices (i.e., ratings), with means μ_1 and μ_2, variances ${\sigma_1}^2$ and ${\sigma_2}^2$, correlation $\rho_{1,2}$, and covariance $\sigma_{1,2} = \rho_{1,2}\sigma_1\sigma_2$. We use quadratic weights and substitute $v(c,\tilde{c}) = (c-\tilde{c})^2/(C-1)^2$, where the denominator drops out. The numerator and denominator in (1) correspond to the expectations of the weight $v(Y_1, Y_2)$ when considering the joint distribution of Y_1 and Y_2 versus independence:

$$\kappa = 1 - \frac{E[v(Y_1,Y_2)]}{E[v(Y_1,Y_2)|\text{independence}]} = 1 - \frac{E\left[(Y_1-Y_2)^2\right]}{E\left[(Y_1-Y_2)^2|\text{independence}\right]}. \tag{2}$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

Table 1. Joint and marginal distributions of ratings in Landis and Koch's multiple sclerosis example with two raters

	$Y_2 = 1$	$Y_2 = 2$	$Y_2 = 3$	$Y_2 = 4$	Total
$Y_1 = 1$	0.0725	0.0435	0.0000	0.0000	0.1159
$Y_1 = 2$	0.0435	0.1594	0.0580	0.0000	0.2609
$Y_1 = 3$	0.0290	0.1884	0.0435	0.0580	0.3188
$Y_1 = 4$	0.0145	0.0290	0.0580	0.2029	0.3043
Total	0.1594	0.4203	0.1594	0.2609	1.0000

Using the definition of variance as the second central moment and writing it in terms of the uncentered moments, we can work out (2) as

$$\begin{aligned}\kappa &= 1 - \frac{\mathrm{Var}\left[Y_1 - Y_2\right] + \left(E\left[Y_1 - Y_2\right]\right)^2}{\mathrm{Var}\left[Y_1 - Y_2|\text{independence}\right] + \left(E\left[Y_1 - Y_2|\text{independence}\right]\right)^2} \\ &= 1 - \frac{\left({\sigma_1}^2 + {\sigma_2}^2 - 2\sigma_{1,2}\right) + \left(\mu_1 - \mu_2\right)^2}{\left({\sigma_1}^2 + {\sigma_2}^2\right) + \left(\mu_1 - \mu_2\right)^2} = \frac{2\sigma_{1,2}}{{\sigma_1}^2 + {\sigma_2}^2 + \left(\mu_1 - \mu_2\right)^2},\end{aligned} \tag{3}$$

showing that Cohen's quadratically weighted kappa coincides with Lin's bivariate concordance correlation coefficient (Lin, 1989). Furthermore, (3) implies that Cohen's quadratically weighted kappa reduces to the Pearson product–moment correlation $\rho_{1,2}$ if we impose equal means and variances, that is, $\mu_1 = \mu_2$ and ${\sigma_1}^2 = {\sigma_2}^2$ (Cohen, 1968).

As an illustration, we consider the second contingency table in Table 1 of Landis and Koch (1977b, p. 161), which compared the multiple sclerosis diagnoses of 69 patients by two neurologists based on four likelihood categories, where "1" is highest and "4" is lowest. Table 1 shows the corresponding joint and marginal distributions.

Cohen's quadratically weighted kappa equals 0.6256. Furthermore, the raters' means, variances, and covariance become

$$\mu_1 = 1 \times 0.1159 + 2 \times 0.2609 + 3 \times 0.3188 + 4 \times 0.3043 = 2.8116,$$

$$\mu_2 = 1 \times 0.1594 + 2 \times 0.4203 + 3 \times 0.1594 + 4 \times 0.2609 = 2.5217,$$

$${\sigma_1}^2 = \left(1 \times 0.1159 + 4 \times 0.2609 + 9 \times 0.3188 + 16 \times 0.3043\right) - \left(2.8116\right)^2 = 0.9935,$$

$${\sigma_2}^2 = \left(1 \times 0.1594 + 4 \times 0.4203 + 9 \times 0.1594 + 16 \times 0.2609\right) - \left(2.5217\right)^2 = 1.0901,$$

and $\sigma_{1,2} = 0.6780$, where we omit the tedious details of the covariance calculation that uses the joint distribution. Lin's bivariate concordance correlation coefficient is

$$\frac{2\sigma_{1,2}}{{\sigma_1}^2 + {\sigma_2}^2 + \left(\mu_1 - \mu_2\right)^2} = \frac{2 \times 0.6780}{0.9935 + 1.0901 + \left(2.8116 - 2.5217\right)^2} = \frac{1.3560}{2.1676} = 0.6256,$$

the same value as Cohen's quadratically weighted kappa.

4. Two interchangeable raters: Scott's quadratically weighted pi

We continue with Scott's pi in the two-rater setting to show how the previous logic for Cohen's quadratically weighted kappa extends to agreement coefficients with interchangeable raters. Later, we extend to Fleiss' kappa with any number of interchangeable raters.

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

4.1. Bivariate concordance correlation after recentering at the raters' grand mean

Expressed in population parameters with C ordered categories, Scott's pi is

$$\pi = 1 - \frac{d_o}{d_e} = 1 - \frac{\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} v(c,\tilde{c}) \Pr(Y_1 = c, Y_2 = \tilde{c})}{\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C} v(c,\tilde{c}) \left(\frac{\Pr(Y_1=c)+\Pr(Y_2=c)}{2}\right)\left(\frac{\Pr(Y_1=\tilde{c})+\Pr(Y_2=\tilde{c})}{2}\right)}, \tag{4}$$

where averaging the marginal category proportions over the two interchangeable raters implies the denominator in (4). Next, incorporating quadratic weights by substituting $v(c,\tilde{c}) = (c-\tilde{c})^2$, recognizing the numerator in (4) as the expectation of the weight $v(Y_1, Y_2)$ when considering the joint distribution of Y_1 and Y_2, and recognizing independence in the denominator yields

$$\pi = 1 - \frac{E\left[(Y_1 - Y_2)^2\right]}{d_e} = 1 - \frac{\left({\sigma_1}^2 + {\sigma_2}^2 - 2\sigma_{1,2}\right) + (\mu_1 - \mu_2)^2}{d_e}, \tag{5}$$

where

$$\begin{aligned}
d_e &= \frac{1}{4}\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C}(c-\tilde{c})^2 \Pr(Y_1 = c)\Pr(Y_1 = \tilde{c}) \\
&\quad + \frac{1}{4}\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C}(c-\tilde{c})^2 \Pr(Y_2 = c)\Pr(Y_2 = \tilde{c}) \\
&\quad + \frac{1}{2}\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C}(c-\tilde{c})^2 \Pr(Y_1 = c)\Pr(Y_2 = \tilde{c}) \\
&= \frac{1}{4}E\left[\left(Y_1 - \tilde{Y}_1\right)^2 | \text{indep.}\right] + \frac{1}{4}E\left[\left(Y_2 - \tilde{Y}_2\right)^2 | \text{indep.}\right] + \frac{1}{2}E\left[(Y_1 - Y_2)^2 | \text{indep.}\right] \\
&= \frac{1}{4}\left(2{\sigma_1}^2 + (\mu_1 - \mu_1)^2\right) + \frac{1}{4}\left(2{\sigma_2}^2 + (\mu_2 - \mu_2)^2\right) + \frac{1}{2}\left({\sigma_1}^2 + {\sigma_2}^2 + (\mu_1 - \mu_2)^2\right) \\
&= {\sigma_1}^2 + {\sigma_2}^2 + \frac{1}{2}(\mu_1 - \mu_2)^2,
\end{aligned} \tag{6}$$

with $\tilde{Y}_1$ and $\tilde{Y}_2$ originating from the same distributions as Y_1 and Y_2, respectively. Substituting (6) into (5) yields a new expression for Scott's pi in terms of the first and second moments:

$$\pi = 1 - \frac{\left({\sigma_1}^2 + {\sigma_2}^2 - 2\sigma_{1,2}\right) + (\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2 + \frac{1}{2}(\mu_1 - \mu_2)^2} = \frac{2\sigma_{1,2} - \frac{1}{2}(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2 + \frac{1}{2}(\mu_1 - \mu_2)^2}. \tag{7}$$

Furthermore, (7) implies that Scott's pi with quadratic weights reduces to the Pearson product–moment correlation $\rho_{1,2}$ if we impose equal means and variances, that is, $\mu_1 = \mu_2$ and ${\sigma_1}^2 = {\sigma_2}^2$.

Comparing expressions (3) and (7) for Cohen's quadratically weighted kappa and the corresponding Scott's pi shows that these coefficients coincide if the two raters share the same mean, $\mu_1 = \mu_2$. This requirement of equal rater means is substantially more relaxed than the typical requirement of raters having the same category proportions for the *unweighted* kappa and pi coefficients. For example, Banerjee et al. (1999, p. 5) noted for nominal categories that "if the two raters are interchangeable, in the sense that the marginal distributions are identical, then Cohen's and Scott's measures are equivalent." In contrast, when the weights are quadratic, having the same mean is already sufficient for interchangeability, making these coefficients equivalent.

Next, we rewrite Scott's pi in (7) as

$$\begin{aligned}
\pi &= \frac{2\left(\sigma_{1,2} - \frac{1}{4}{\mu_1}^2 - \frac{1}{4}{\mu_2}^2 + \frac{1}{2}\mu_1\mu_2\right)}{\left({\sigma_1}^2 + \frac{3}{4}{\mu_1}^2 - \frac{1}{4}{\mu_2}^2 - \frac{1}{2}\mu_1\mu_2\right) + \left({\sigma_2}^2 - \frac{1}{4}{\mu_1}^2 + \frac{3}{4}{\mu_2}^2 - \frac{1}{2}\mu_1\mu_2\right)} \\
&= \frac{2\left(\sigma_{1,2} + \mu_1\mu_2 - \left(\frac{\mu_1+\mu_2}{2}\right)^2\right)}{\left({\sigma_1}^2 + {\mu_1}^2 - \left(\frac{\mu_1+\mu_2}{2}\right)^2\right) + \left({\sigma_2}^2 + {\mu_2}^2 - \left(\frac{\mu_1+\mu_2}{2}\right)^2\right)} = \frac{2\tilde{\sigma}_{1,2}}{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2},
\end{aligned} \tag{8}$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

where $\tilde{\sigma}_{1,2} = \sigma_{1,2} + \mu_1\mu_2 - \overline{\mu}^2$ is the covariance after recentering the rater means, μ_1 and μ_2, at the grand mean, $\overline{\mu} = (\mu_1 + \mu_2)/2$, and $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$ are the corresponding recentered variances. Thus, we may interpret Cohen's quadratically weighted kappa in (3) and Scott's quadratically weighted pi in (7) as concordance correlation coefficients, where the interchangeable raters in Scott's pi imply recentering the rater means and (co)variances at the grand mean.

4.2. Pearson correlation and regression slope after concatenating the ratings

Another interpretation of Scott's quadratically weighted pi is that it equals the Pearson product–moment correlation when the order of the two raters is random, with each of the two possible rater pairs, (1,2) and (2,1), occurring with 50% probability. Since the rater order is random, the two raters in the pair become random and effectively interchangeable with the same mean and variance. Following earlier logic, using equal means and variances, and denoting the two random raters as J_1 and J_2, we indeed obtain that Scott's pi is the corresponding Pearson correlation:

$$\begin{aligned}\pi &= 1 - \frac{E\left[(Y_{J_1} - Y_{J_2})^2\right]}{E\left[(Y_{J_1} - Y_{J_2})^2 | \text{independence}\right]} = 1 - \frac{\text{Var}[Y_{J_1}] + \text{Var}[Y_{J_2}] - 2\,\text{Cov}[Y_{J_1}, Y_{J_2}]}{\text{Var}[Y_{J_1}] + \text{Var}[Y_{J_2}]} \\ &= 1 - \frac{2\,\text{Var}[Y_{J_1}] - 2\,\text{Cov}[Y_{J_1}, Y_{J_2}]}{2\,\text{Var}[Y_{J_1}]} = \frac{\text{Cov}[Y_{J_1}, Y_{J_2}]}{\text{Var}[Y_{J_1}]} = \frac{\text{Cov}[Y_{J_1}, Y_{J_2}]}{\sqrt{\text{Var}[Y_{J_1}]\,\text{Var}[Y_{J_2}]}} = \text{Corr}[Y_{J_1}, Y_{J_2}]. \end{aligned} \tag{9}$$

In practice, researchers can compute Scott's quadratically weighted pi by computing the Pearson correlation after concatenating the data, that is, combining the ratings from the two possible rater pairs, (1,2) and (2,1), in one extended data set with twice as many ratings. This procedure starts from the original contingency table, containing the rating frequencies of the two raters, and adds the transpose of this contingency table, making the combined contingency table symmetric. As the raters always become interchangeable after combining the two possible rater orders, it does not matter whether the same two raters rated all items or raters varied across items. However, the variance (and standard error) of this Pearson correlation is invalid because the ratings of the flipped rater pair are a direct consequence of the original ratings, violating the requirement of being drawn independently from the same data-generating process. Gwet (2014, p. 141) and Andres and Hernandez (2025) provided formulas to compute the correct variance of Scott's pi for fixed raters. Furthermore, Gwet (2014) proposed an additive jackknife variance component if the raters are random (i.e., the participating raters are not of central interest and instead drawn from a larger population they are supposed to represent).

Next, we note that (9) implies $\pi = \beta$, where β is the slope in $Y_{J_2} = \alpha + \beta Y_{J_1} + \varepsilon$. The reason is that the slope coefficient in a simple linear regression equals the covariance between the independent variable and the dependent variable divided by the variance of the independent variable, that is, $\beta = \text{Cov}[Y_{J_1}, Y_{J_2}]/\text{Var}[Y_{J_1}]$ (e.g., Wooldridge, 2012, p. 29). Thus, analogous to before, researchers can compute Scott's quadratically weighted pi by estimating the slope coefficient in a simple linear regression for the concatenated ratings of rater pairs (1,2) and (2,1). However, standard regression theory no longer helps obtain the corresponding variance.

Moving from the left to the right panel in Figure 2 shows how the concatenation works in the multiple sclerosis example of Landis and Koch (1977b). For each cell, we sum the original rating frequency of the rater pair (1,2) and the diagonal-mirrored frequency of the flipped rater pair (2,1). Fleiss' kappa equals the Pearson correlation and regression slope in the figure's right panel.

4.3. Linear function of Cohen's quadratically weighted kappa

We write Scott's quadratically weighted pi in (7) as a linear function of Cohen's quadratically weighted kappa in (3), conditional on the rater means and variances (Appendix A):

$$\pi = \kappa - \left(\frac{W}{2+W}\right)(1-\kappa) \leq \kappa, \tag{10}$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

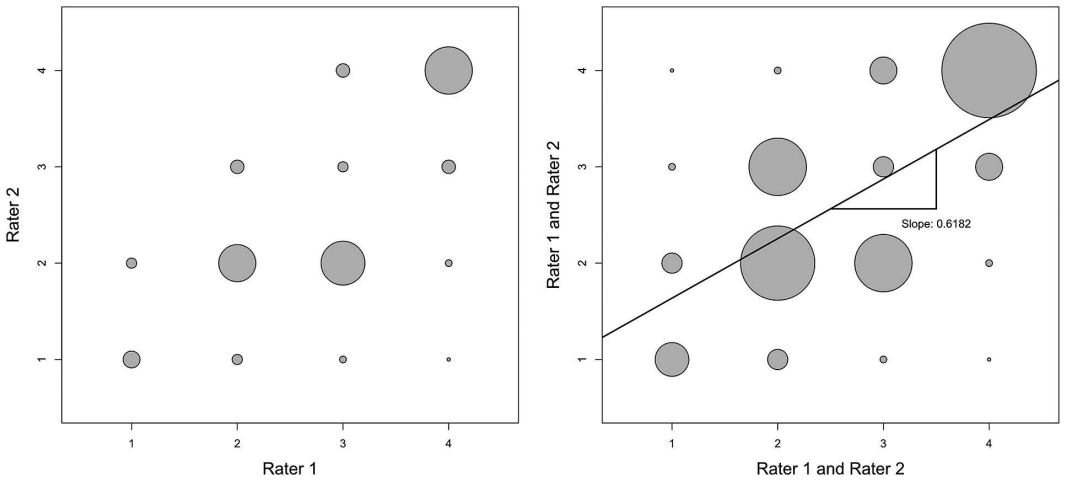


Figure 2. After concatenation, Scott's quadratically weighted pi equals the Pearson correlation and regression slope of 0.6182 in Landis and Koch's two-rater multiple sclerosis example. Scatter plot of original frequencies (left) and combined frequencies with regression line (right).

where

$$W = \frac{(\mu_1 - \mu_2)^2}{\sigma_1{}^2 + \sigma_2{}^2} \geq 0 \tag{11}$$

is the squared difference in rater means divided by the corresponding variance under independence. Its sample counterpart is related to the Wald (or squared t) statistic to test for equal means while allowing for unequal variances. Equation (10) shows that Scott's pi cannot exceed Cohen's kappa, so the latter coefficient provides an upper bound for the former. Whereas Warrens (2010a) proved this result for nominal weights, we demonstrate it for quadratic weights.

Although Scott's quadratically weighted pi is a linear function of Cohen's quadratically weighted kappa *conditional on* the rater means and variances, the relationship between the variance estimators of these coefficients remains complex. The reason is that both the coefficients and rater moments are functions of the data and interact nonlinearly in (10) and (11).

4.4. Numerical example

Continuing with the multiple sclerosis example of Landis and Koch (1977b, p. 161), we use (7) to compute Scott's quadratically weighted pi as

$$\pi = \frac{2 \times 0.6780 - \frac{1}{2} \times (2.8116 - 2.5217)^2}{0.9935 + 1.0901 + \frac{1}{2} \times (2.8116 - 2.5217)^2} = \frac{1.3140}{2.1256} = 0.6182.$$

Next, we use (8) to compute Scott's pi as Lin's concordance correlation coefficient after recentering at the raters' grand mean, which is $\overline{\mu} = (2.8116 + 2.5217)/2 = 2.6667$:

$$\begin{aligned}
\tilde{\sigma}_{1,2} &= \sigma_{1,2} + \mu_1\mu_2 - \overline{\mu}^2 = 0.6780 + 2.8116 \times 2.5217 - (2.6667)^2 = 0.6570, \\
\tilde{\sigma}_1^2 &= \sigma_1{}^2 + \mu_1{}^2 - \overline{\mu}^2 = 0.9935 + (2.8116)^2 - (2.6667)^2 = 1.7876, \\
\tilde{\sigma}_2^2 &= \sigma_2{}^2 + \mu_2{}^2 - \overline{\mu}^2 = 1.0901 + (2.5217)^2 - (2.6667)^2 = 0.3381, \\
\pi &= \frac{2\tilde{\sigma}_{1,2}}{\tilde{\sigma}_1^2 + \tilde{\sigma}_2^2} = \frac{2 \times 0.6570}{1.7876 + 0.3381} = \frac{1.3140}{2.1256} = 0.6182,
\end{aligned}$$

yielding the same coefficient value.

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

We proceed by computing Scott's quadratically weighted pi as the Pearson product–moment correlation and the regression slope after concatenating the ratings of the original rater pair (1,2) and flipped rater pair (2,1), that is, adding the transposed contingency table to the original one. The correlation and slope coefficients become the covariance divided by the variance (where the two data columns share the same mean and variance after concatenation). The R-code below starts from Landis and Koch's 4×4 contingency table and avoids translating this contingency table into item-by-rater tables of ratings. It again yields a value of 0.6182:

```
y <- matrix(c(5,3,2,1,3,11,13,2,0,4,3,4,0,0,4,14),nrow=4)
y <- (y+t(y))/(2*sum(y))
mean <- sum(1:nrow(y)*rowSums(y))
var <- sum((1:nrow(y))^2*rowSums(y))-mean^2
cov <- sum(outer(1:nrow(y),1:nrow(y))*y)-mean^2
cov/var
[1] 0.6181818
```

Furthermore, Scott's quadratically weighted pi is a linear function of Cohen's quadratically weighted kappa (and vice versa), with the intercept and slope solely depending on the squared difference in means divided by the corresponding variance under independence:

$$W = \frac{(\mu_1 - \mu_2)^2}{\sigma_1{}^2 + \sigma_2{}^2} = \frac{(2.8116 - 2.5217)^2}{0.9935 + 1.0901} = \frac{0.0840}{2.0836} = 0.0403.$$

Substituting into (10) yields

$$\frac{W}{2+W} = \frac{0.0403}{2.0403} = 0.0198,$$
$$\pi = 0.6182 = 0.6256 - 0.0198 \times (1 - 0.6256) = \kappa - \left(\frac{W}{2+W}\right)(1-\kappa). \tag{12}$$

5. Any noninterchangeable raters: Conger's quadratically weighted kappa

To explain how the previous ideas extend to multirater coefficients with more than two raters, we show that Conger's quadratically weighted kappa equals Lin's generalized concordance correlation coefficient and reduces to the average Pearson product–moment correlation across all rater pairs if raters share the same mean and variance. This proof is shorter and less complex than found elsewhere in the literature (Andres & Hernandez, 2020; Schuster & Smith, 2005; Warrens, 2012a); the result for the average Pearson correlation is new. Later, we extend to the more complex analysis of Fleiss' quadratically weighted kappa with interchangeable raters.

We denote Conger's quadratically weighted kappa for R raters as κ_C to distinguish it from Cohen's κ for two raters. In Conger's kappa, both the observed disagreement, d_o, and expected disagreement, d_e, are averages (or equivalently, sums) across all pairs of different raters $(r, \tilde{r})$ with $\tilde{r} > r$ (Conger, 1980):

$$\kappa_C = 1 - \frac{d_o}{d_e} = 1 - \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left[\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C}(c-\tilde{c})^2 \Pr\left(Y_r = c, Y_{\tilde{r}} = \tilde{c}\right)\right]}{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left[\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C}(c-\tilde{c})^2 \Pr\left(Y_r = c\right)\Pr\left(Y_{\tilde{r}} = \tilde{c}\right)\right]}. \tag{13}$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

Recognizing the expectations in square brackets, with independence in the denominator, yields

$$\begin{aligned}\kappa_C &= 1 - \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r} E\left[\left(Y_r - Y_{\tilde{r}}\right)^2\right]}{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r} E\left[\left(Y_r - Y_{\tilde{r}}\right)^2 | \text{independence}\right]} \\ &= 1 - \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r} \text{Var}\left[Y_r - Y_{\tilde{r}}\right] + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(E\left[Y_r - Y_{\tilde{r}}\right]\right)^2}{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r} \text{Var}\left[Y_r - Y_{\tilde{r}} | \text{independence}\right] + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(E\left[Y_r - Y_{\tilde{r}} | \text{independence}\right]\right)^2} \\ &= 1 - \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2 - 2\sigma_{r,\tilde{r}}\right) + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2\right) + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}. \end{aligned} \tag{14}$$

Next, we use that $\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2\right) = (R-1)\sum_{r=1}^{R}{\sigma_r}^2$ and simplify (14) to obtain Lin's generalized concordance correlation coefficient, suitable for any number of raters (Andres & Hernandez, 2020; Barnhart et al., 2002; Carrasco & Jover, 2003; Lin, 1989, 2000):

$$\kappa_C = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}}}{(R-1)\sum_{r=1}^{R}{\sigma_r}^2 + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}. \tag{15}$$

We note that (15) reduces to Cohen's quadratically weighted kappa and the bivariate concordance correlation coefficient in (3) if there are only two raters, with $R = 2$.

Conger's kappa in (15) becomes the average Pearson product–moment correlation across all rater pairs if raters share the same mean and the same variance, σ^2:

$$\kappa_C \mid \text{same mean and variance} = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\rho_{r,\tilde{r}}\sigma^2}{(R-1)\left(R\sigma^2\right)} = \frac{\sigma^2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\rho_{r,\tilde{r}}}{\frac{\sigma^2 R(R-1)}{2}} = \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\rho_{r,\tilde{r}}}{\binom{R}{2}}. \tag{16}$$

Finally, we express (15) in matrix and vector notation for convenient coding in matrix programming languages:

$$\kappa_C = \frac{\iota'\Sigma\iota - \text{tr}\left(\Sigma\right)}{(R-1)\,\text{tr}\left(\Sigma\right) + R^2\left(\overline{\mu^2} - \left(\overline{\mu}\right)^2\right)}, \tag{17}$$

where $'$ denotes the transpose, ι is the R-dimensional vector of ones, $\Sigma = \left(\sigma_{r,\tilde{r}}\right)$ is the covariance matrix, "tr" is the trace (i.e., sum of diagonal elements), $\overline{\mu} = \iota'\mu/R$, and $\overline{\mu^2} = \iota'\mu^2/R$, with μ and μ^2 being R-dimensional vectors containing the rater means and squared rater means. Comparing (17) with the matrix expressions in Almehrizi and Emam (2023) reveals identical numerators and related denominators. Their two-factor intraclass correlation coefficient for absolute agreement (labeled $\rho_{(A,1)}$) converges to Conger's quadratically weighted kappa in (17), as their only deviating term converges to $(R-1)\,\text{tr}\left(\Sigma\right)$. Furthermore, the term $R^2\left(\overline{\mu^2} - \left(\overline{\mu}\right)^2\right)$ is absent in their two-factor intraclass correlation coefficient for relative consistency (labeled $\rho_{(C,1)}$). Hence, this coefficient coincides with Conger's kappa in (17) if all raters share the same mean.

As an illustration, we consider Table A.3 of Gwet (2014, p. 371), showing the five-point color intensity ratings of 29 stickleback fishes by four experienced raters. Based on maximum likelihood estimation, these raters have the following mean vector and covariance matrix:

$$\mu = \begin{pmatrix} 2.8276 \\ 2.6207 \\ 2.7931 \\ 2.8276 \end{pmatrix},$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

$$\Sigma = \begin{pmatrix} 2.3496 & 1.6932 & 1.7574 & 1.6254 \\ 1.6932 & 2.0285 & 1.5767 & 1.7967 \\ 1.7574 & 1.5767 & 2.3020 & 2.0678 \\ 1.6254 & 1.7967 & 2.0678 & 2.8323 \end{pmatrix},$$

implying that $\iota'\Sigma\iota = 30.5470$, $\mathrm{tr}(\Sigma) = 9.5125$, $\overline{\mu} = 2.7672$, and $\overline{\mu^2} = 7.6650$. Plugging these numbers into (17) yields the value of Conger's quadratically weighted kappa and the generalized concordance correlation coefficient, which is also the value reported in Gwet (2014, p. 158):

$$\kappa_C = \frac{30.5470 - 9.5125}{3 \times 9.5125 + 16 \times \left(7.6650 - (2.7672)^2\right)} = \frac{21.0345}{28.6552} = 0.7341.$$

Letting y denote the 29 × 4 item-by-rater data matrix from Gwet (2014, p. 371), then coding up (17) in R is straightforward and yields the same coefficient value:

```
R <- ncol(y)
mu <- colMeans(y)
Sigma <- cov(y)*((nrow(y)-1)/nrow(y)) # ML estimation
num <- sum(Sigma)-sum(diag(Sigma))
den <- (R-1)*sum(diag(Sigma))+(R^2)*(mean(mu^2)-(mean(mu))^2)
num/den
[1] 0.7340554
```

In all remaining programming codes, we will refer to the data matrix as *y* and skip the calculations of *R* (i.e., number of raters), μ, and Σ because these calculations remain identical.

6. Any interchangeable raters: Fleiss' quadratically weighted kappa

We turn to the coefficient of primary interest, Fleiss' quadratically weighted kappa with any number of interchangeable raters, denoted as κ_F. These results are new.

6.1. Generalized concordance correlation after recentering at the raters' grand mean

Fleiss' kappa computes the observed disagreement, d_o, in the same way as Conger's: It takes the corresponding average over all pairs of different raters. However, unlike Conger's kappa, Fleiss' chance correction aggregates the category proportions across raters and uses these average proportions to compute its expected disagreement, d_e (Fleiss, 1971):

$$\kappa_F = 1 - \frac{d_o}{d_e} = 1 - \frac{\dfrac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left[\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C}(c-\tilde{c})^2 \Pr\left(Y_r = c, Y_{\tilde{r}} = \tilde{c}\right)\right]}{\binom{R}{2}}}{\sum_{c=1}^{C}\sum_{\tilde{c}=1}^{C}(c-\tilde{c})^2 \left(\frac{\sum_{r=1}^{R}\Pr(Y_r=c)}{R}\right)\left(\frac{\sum_{\tilde{r}=1}^{R}\Pr(Y_{\tilde{r}}=\tilde{c})}{R}\right)}. \tag{18}$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

Using that $R^2/\binom{R}{2} = 2R/(R-1)$ and recognizing the expectations, we can write (18) as

$$\begin{aligned}
\kappa_F &= 1 - \frac{2R}{R-1} \times \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r} E\left[(Y_r - Y_{\tilde{r}})^2\right]}{\sum_{r=1}^{R}\sum_{\tilde{r}=1}^{R} E\left[(Y_r - Y_{\tilde{r}})^2 | \text{independence}\right]} \\
&= 1 - \frac{2R}{R-1} \times \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r} \text{Var}\left[Y_r - Y_{\tilde{r}}\right] + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(E\left[Y_r - Y_{\tilde{r}}\right]\right)^2}{\sum_{r=1}^{R}\sum_{\tilde{r}=1}^{R} \text{Var}\left[Y_r - Y_{\tilde{r}} | \text{indep.}\right] + \sum_{r=1}^{R}\sum_{\tilde{r}=1}^{R}\left(E\left[Y_r - Y_{\tilde{r}} | \text{indep.}\right]\right)^2} \\
&= 1 - \frac{2R}{R-1} \times \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2 - 2\sigma_{r,\tilde{r}}\right) + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}{\sum_{r=1}^{R}\sum_{\tilde{r}=1}^{R}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2\right) + \sum_{r=1}^{R}\sum_{\tilde{r}=1}^{R}\left(\mu_r - \mu_{\tilde{r}}\right)^2}.
\end{aligned} \tag{19}$$

Next, we break up the summations in the denominator into $r = \tilde{r}$, $r < \tilde{r}$, and symmetric $r > \tilde{r}$:

$$\kappa_F = 1 - \frac{R}{R-1} \times \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2 - 2\sigma_{r,\tilde{r}}\right) + 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}{\left[2\sum_{r=1}^{R}{\sigma_r}^2 + 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2\right)\right] + 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}. \tag{20}$$

Finally, we divide the numerator and denominator by two, use that $\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2\right) = (R-1)\sum_{r=1}^{R}{\sigma_r}^2$, and simplify to obtain

$$\begin{aligned}
\kappa_F &= 1 - \frac{R}{R-1} \times \frac{(R-1)\sum_{r=1}^{R}{\sigma_r}^2 - 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}} + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}{R\sum_{r=1}^{R}{\sigma_r}^2 + \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2} \\
&= 1 - \frac{R(R-1)\sum_{r=1}^{R}{\sigma_r}^2 - 2R\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}} + R\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}{R(R-1)\sum_{r=1}^{R}{\sigma_r}^2 + (R-1)\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2} \\
&= \frac{2R\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}} - \sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}{R(R-1)\sum_{r=1}^{R}{\sigma_r}^2 + (R-1)\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2} \\
&= \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}} - \frac{1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}{(R-1)\sum_{r=1}^{R}{\sigma_r}^2 + \frac{R-1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\mu_r - \mu_{\tilde{r}}\right)^2}.
\end{aligned} \tag{21}$$

Using the same matrix and vector notation as before, this expression for Fleiss' quadratically weighted kappa becomes

$$\kappa_F = \frac{\iota'\Sigma\iota - \text{tr}(\Sigma) - R\left(\overline{\mu^2} - (\overline{\mu})^2\right)}{(R-1)\,\text{tr}(\Sigma) + R(R-1)\left(\overline{\mu^2} - (\overline{\mu})^2\right)}. \tag{22}$$

We note that (21) reduces to Scott's quadratically weighted pi in (7) if there are only two raters, with $R = 2$. Furthermore, it becomes the average Pearson product–moment correlation across all rater pairs if raters share the same mean and the same variance, σ^2:

$$\kappa_F \mid \text{same mean and variance} = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\rho_{r,\tilde{r}}\sigma^2}{(R-1)(R\sigma^2)} = \frac{\sigma^2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\rho_{r,\tilde{r}}}{\frac{\sigma^2 R(R-1)}{2}} = \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\rho_{r,\tilde{r}}}{\binom{R}{2}}. \tag{23}$$

Comparing expressions (15) and (21) for Conger's and Fleiss' quadratically weighted kappas shows that these coefficients coincide if all raters share the same mean, which is a much less stringent condition than all raters having identical category proportions (i.e., the same marginal distributions).

Next, we show that Fleiss' quadratically weighted kappa in (21) equals Lin's generalized concordance correlation coefficient after recentering the rater means and covariance matrix at the raters' grand mean (Appendix B):

$$\kappa_F = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\tilde{\sigma}_{r,\tilde{r}}}{(R-1)\sum_{r=1}^{R}\tilde{\sigma}_r^2} = \frac{\iota'\tilde{\Sigma}\iota - \text{tr}\left(\tilde{\Sigma}\right)}{(R-1)\,\text{tr}\left(\tilde{\Sigma}\right)}, \tag{24}$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

where $\tilde{\sigma}_{r,\tilde{r}} = \sigma_{r,\tilde{r}} + \mu_r\mu_{\tilde{r}} - \overline{\mu}^2$ denotes the covariances after recentering at $\overline{\mu} = \sum_{r=1}^{R}\mu_r/R$ and $\tilde{\sigma}_r^2 = {\sigma_r}^2 + {\mu_r}^2 - \overline{\mu}^2$ denotes the recentered variances. Using the same matrix and vector notation as before, the recentered covariance matrix becomes $\tilde{\Sigma} = \Sigma + \mu\mu' - \overline{\mu}^2$. This result generalizes an earlier result, obtained in (8) for Scott's pi with two raters, to any number of raters.

6.2. Pearson correlation and regression slope after concatenating the ratings

Fleiss' quadratically weighted kappa is the multirater generalization of the corresponding Scott's pi. As Scott's pi equals the Pearson product–moment correlation between the ratings from two random, different raters, Fleiss' kappa has the same interpretation. Treating the two raters in the pair as random makes these raters interchangeable with the same mean and variance, so the same proof as in (9) applies. In practice, researchers can compute Fleiss' kappa with quadratic weights by appropriately concatenating the data and then computing the Pearson correlation from these concatenated ratings. The concatenation procedure generalizes the approach for Scott's pi. It combines the ratings from all possible rater pairs (including the same pair after flipping the two raters) in a new data set, where the first column captures the first rater in the pair, and the second column captures the second rater. For example, for three raters, the two data columns combine (i.e., concatenate) the ratings from the six rater pairs (1,2), (2,1), (1,3), (3,1), (2,3), and (3,2).

Furthermore, analogous to before, researchers can compute the value of Fleiss' quadratically weighted kappa by estimating the slope coefficient in a simple linear regression, where the two data columns with concatenated ratings serve as dependent and independent variables (or vice versa). However, standard statistical theory no longer helps obtain the variances, as the concatenated ratings are no independent draws from the data-generating process. We refer to other literature sources to get the correct variance of Fleiss' kappa (Andres & Hernandez, 2025; Gwet, 2014; Schouten, 1980, 1982; Vanbelle, 2019). Furthermore, a high-quality R package to compute agreement coefficients and their variances is *irrCAC*.

6.3. Linear function of Conger's quadratically weighted kappa

Generalizing the result in (10) for two raters to any number of raters, we write Fleiss' quadratically weighted kappa in (21) as a linear function of Conger's quadratically weighted kappa in (15), conditional on the rater means and variances (Appendix C):

$$\kappa_F = \kappa_C - \left(\frac{W}{R + (R-1)\,W}\right)(1 - \kappa_C) \leq \kappa_C, \tag{25}$$

where

$$W = \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r - \mu_{\tilde{r}})^2}{(R-1)\sum_{r=1}^{R}{\sigma_r}^2} = \frac{R^2\left(\overline{\mu^2} - (\overline{\mu})^2\right)}{(R-1)\,\mathrm{tr}\,(\Sigma)} \geq 0 \tag{26}$$

is the sum of squared differences in means across all rater pairs divided by the corresponding sum of variances under independence. We recall $\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left({\sigma_r}^2 + {\sigma_{\tilde{r}}}^2\right) = (R-1)\sum_{r=1}^{R}{\sigma_r}^2$. Equation (25) shows that Fleiss' quadratically weighted kappa cannot exceed Conger's, implying that the latter coefficient provides an upper bound for the former. Whereas Warrens (2010b) proved this result for nominal weights, we extend it to quadratic weights. Furthermore, (25) provides an alternative proof that Conger's and Fleiss' quadratically weighted kappas converge to each other as the number of raters, R, tends to infinity (Conger, 1980; Gwet, 2014).

Although Fleiss' quadratically weighted kappa is a linear function of Conger's quadratically weighted kappa *conditional on* the rater means and variances, the relationship between the variance estimators of these kappas remains complex. The reason is that both the kappas and rater moments are functions of the data and interact nonlinearly in (25) and (26).

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

6.4. Numerical example

Continuing with the four-rater stickleback fish color example of Gwet (2014, p. 371) and using our earlier calculations of $\iota'\Sigma\iota = 30.5470$, $\text{tr}(\Sigma) = 9.5125$, $\overline{\mu} = 2.7672$, and $\overline{\mu^2} = 7.6650$, we plug these numbers into (22) to compute Fleiss' quadratically weighted kappa as

$$\kappa_F = \frac{30.5470 - 9.5125 - 4 \times \left(7.6650 - (2.7672)^2\right)}{3 \times 9.5125 + 4 \times 3 \times \left(7.6650 - (2.7672)^2\right)} = \frac{21.0051}{28.6257} = 0.7338,$$

which is also the value reported in Gwet (2014, p. 158). As the four raters have almost the same mean rating (only the second rater deviates somewhat), it is unsurprising that the values of Conger's and Fleiss' kappas are virtually identical (0.7341 vs. 0.7338). Indeed, we identified equal rater means as a sufficient condition for these two coefficients, with quadratic weights, to coincide.

We may also compute Fleiss' quadratically weighted kappa as Lins' generalized concordance correlation coefficient after recentering at the raters' grand mean, resulting in the transformed covariance matrix:

$$\tilde{\Sigma} = (\tilde{\sigma}_{r,\tilde{r}}) = \left(\sigma_{r,\tilde{r}} + \mu_r\mu_{\tilde{r}} - \overline{\mu}^2\right) = \Sigma + \mu\mu' - \overline{\mu}^2 = \begin{pmatrix} 2.6872 & 1.4458 & 1.9975 & 1.9631 \\ 1.4458 & 1.2389 & 1.2389 & 1.5493 \\ 1.9975 & 1.2389 & 2.4458 & 2.3079 \\ 1.9631 & 1.5493 & 2.3079 & 3.1700 \end{pmatrix},$$

where $\iota'\tilde{\Sigma}\iota = 30.5470$ and $\text{tr}\left(\tilde{\Sigma}\right) = 9.5419$. Continuing with matrix and vector notation and substituting into (24) yields the same coefficient value as above:

$$\kappa_F = \frac{\iota'\tilde{\Sigma}\iota - \text{tr}\left(\tilde{\Sigma}\right)}{(R-1)\,\text{tr}\left(\tilde{\Sigma}\right)} = \frac{30.5470 - 9.5419}{3 \times 9.5419} = \frac{21.0051}{28.6257} = 0.7338.$$

Coding up Fleiss' quadratically weighted kappa using the rater means and covariance matrix via (22) and as a generalized concordance correlation via (24) yields the same results:

```
num <- sum(Sigma)-sum(diag(Sigma))-R*(mean(mu^2)-(mean(mu))^2)
den <- (R-1)*sum(diag(Sigma))+R*(R-1)*(mean(mu^2)-(mean(mu))^2)
num/den
[1] 0.7337819
```

and

```
Sigma_tilde <- Sigma+outer(mu,mu)-(mean(mu))^2
num <- sum(Sigma_tilde)-sum(diag(Sigma_tilde))
den <- (R-1)*sum(diag(Sigma_tilde))
num/den
[1] 0.7337819
```

Furthermore, Fleiss' quadratically weighted kappa is a linear function of Conger's quadratically weighted kappa (and vice versa). Plugging the numbers into (26) yields

$$W = \frac{R^2\left(\overline{\mu^2} - (\overline{\mu})^2\right)}{(R-1)\,\text{tr}(\Sigma)} = \frac{16 \times \left(7.6650 - (2.7672)^2\right)}{3 \times 9.5125} = \frac{0.1217}{28.5375} = 0.00412,$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

and substituting into (25) yields

$$\frac{W}{R+(R-1)W} = \frac{0.00412}{4+3\times 0.00412} = \frac{0.00412}{4.01237} = 0.00103,$$

$$\kappa_F = 0.7338 = 0.7341 - 0.00103 \times 0.2659 = \kappa_C - \left(\frac{W}{R+(R-1)W}\right)(1-\kappa_C).$$

Coding up this linear connection in R, starting from $\kappa_C = 0.7340554$, yields the correct κ_F:

```
kappa_C <- 0.7340554
W <- (R^2)*(mean(mu^2)-(mean(mu))^2)/((R-1)*sum(diag(Sigma)))
kappa_C-(W/(R+(R-1)*W))*(1-kappa_C)
[1] 0.733782
```

Finally, we show how researchers can compute Fleiss' quadratically weighted kappa as either the Pearson product–moment correlation or the regression slope after appropriately concatenating the ratings into two columns. Since we have four raters, we combine the ratings from 12 rater pairs: (1,2), (1,3), (1,4), (2,3), (2,4), (3,4), (2,1), (3,1), (4,1), (3,2), (4,2), and (4,3), where we may list these pairs in any possible order. The R code for the Pearson correlation of the concatenated ratings is as follows:

```
y1 <- c(y[,1],y[,1],y[,1],y[,2],y[,2],y[,3])
y2 <- c(y[,2],y[,3],y[,4],y[,3],y[,4],y[,4])
cor(c(y1,y2),c(y2,y1))
[1] 0.7337819
```

We can also run the corresponding simple linear regression and inspect the slope:

```
unname(coef(lm(c(y1,y2)~c(y2,y1)))[2])
[1] 0.7337819
```

7. Discussion

We provided a review of distribution-based agreement coefficients with quadratic weights. This extensive class includes the most frequently used and cited coefficients in the literature: Cohen's kappa for two raters and Fleiss' kappa for more than two raters. Furthermore, quadratic weights are the most common in empirical studies (Vanbelle, 2016; Warrens, 2014). These weights possess the attractive property that many distribution-based coefficients become correlation coefficients, facilitating interpretation. Figure 3 provides an overview. In addition to four classic agreement coefficients (i.e., Scott's pi, Cohen's, Conger's, and Fleiss' kappas), the correlation interpretations extend to other multirater coefficients. Light's kappa is the average of all pairwise Cohen's kappas. Hubert's kappa defines rater (dis)agreements by the *joint* outcome of all R raters; it becomes Conger's kappa if the R-rater weight is the corresponding average weight across all rater pairs, which is a well-documented choice (Andres & Hernandez, 2020; Mielke et al., 2007, 2009; Schuster & Smith, 2005; Warrens, 2012a). Krippendorff's alpha and the recently introduced uniform prior coefficient involve traditional versus Bayesian small-sample corrections and converge to Fleiss' kappa. The interpretations for interchangeable raters in the right half of Figure 3 are new. Some of these results extend existing interpretations to interchangeable raters. Other results are unique, such as the Pearson correlation and regression slope after concatenating all rater pairs (including

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

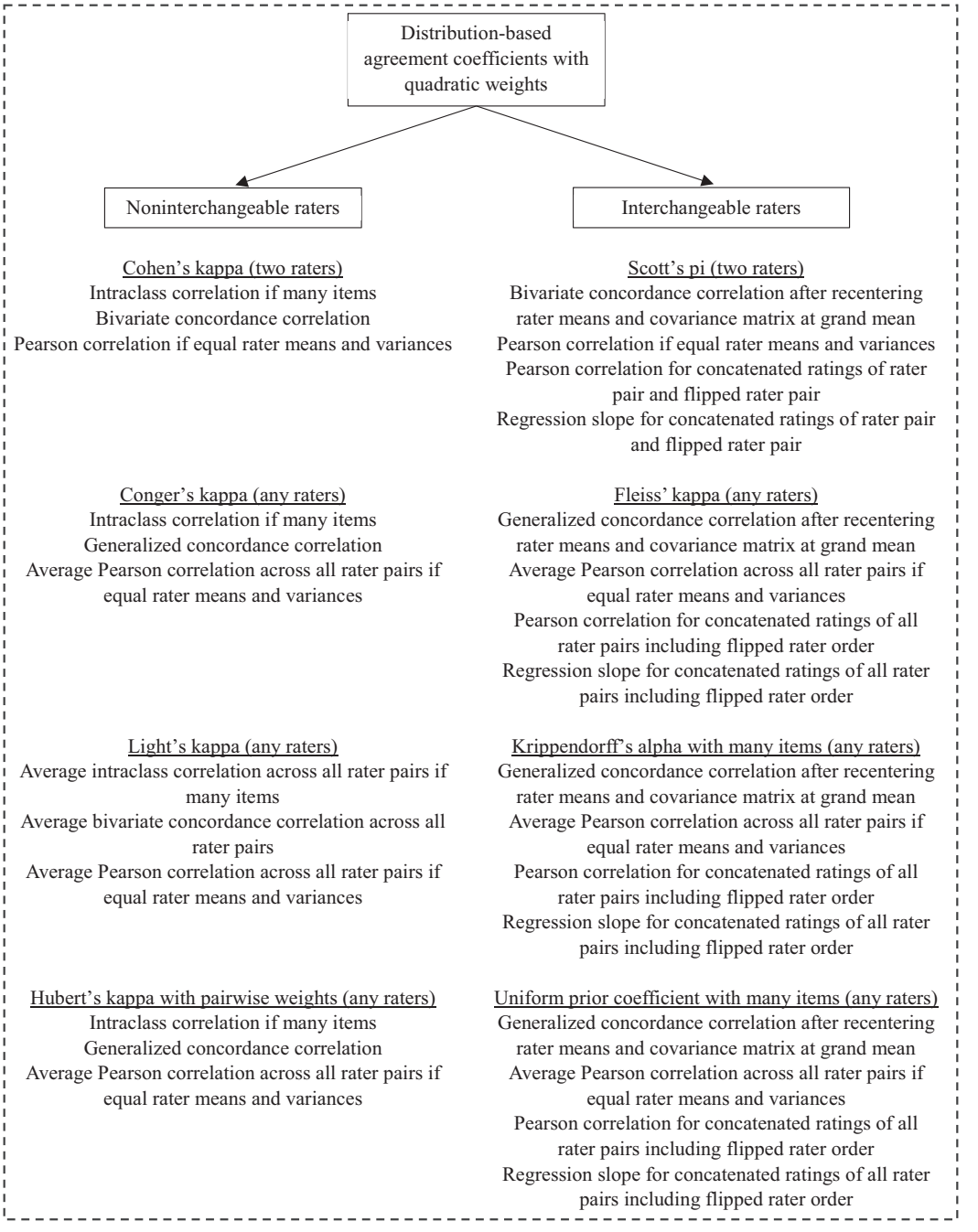


Figure 3. Correlation interpretations of distribution-based agreement coefficients with quadratic weights.

reversed rater order). Most interpretations for noninterchangeable raters in the left half of Figure 3 already exist in the literature.

Table 2 shows the connections between quadratically weighted agreement coefficients that assume noninterchangeable raters (Cohen's and Conger's kappas) versus interchangeable raters (Scott's pi and Fleiss' kappa). Interchangeability implies that raters are treated *identically* in the coefficient's chance

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

Table 2. Expressions and connections of distribution-based agreement coefficients with quadratic weights

	Noninterchangeable raters	Interchangeable raters
Two raters	Cohen's kappa	Scott's pi
Expression:	$\kappa = \frac{2\sigma_{1,2}}{{\sigma_1}^2+{\sigma_2}^2+(\mu_1-\mu_2)^2}$	$\pi = \frac{2\sigma_{1,2}-\frac{1}{2}(\mu_1-\mu_2)^2}{{\sigma_1}^2+{\sigma_2}^2+\frac{1}{2}(\mu_1-\mu_2)^2}$
Connection:	$\kappa = \pi + \left(\frac{W}{2+2W}\right)(1-\pi)$	$\pi = \kappa - \left(\frac{W}{2+W}\right)(1-\kappa)$
	where $W = \frac{(\mu_1-\mu_2)^2}{{\sigma_1}^2+{\sigma_2}^2} \geq 0$	where $W = \frac{(\mu_1-\mu_2)^2}{{\sigma_1}^2+{\sigma_2}^2} \geq 0$
	$\kappa \geq \pi$	$\pi \leq \kappa$
	$\kappa = \pi$ if $\mu_1 = \mu_2$	$\pi = \kappa$ if $\mu_1 = \mu_2$
Any raters	Conger's kappa	Fleiss' kappa
Expression:	$\kappa_C = \frac{\iota'\Sigma\iota-\mathrm{tr}(\Sigma)}{(R-1)\mathrm{tr}(\Sigma)+R^2\left(\overline{\mu^2}-(\overline{\mu})^2\right)}$	$\kappa_F = \frac{\iota'\Sigma\iota-\mathrm{tr}(\Sigma)-R\left(\overline{\mu^2}-(\overline{\mu})^2\right)}{(R-1)\mathrm{tr}(\Sigma)+R(R-1)\left(\overline{\mu^2}-(\overline{\mu})^2\right)}$
Connection:	$\kappa_C = \kappa_F + \left(\frac{W}{R(1+W)}\right)(1-\kappa_F)$	$\kappa_F = \kappa_C - \left(\frac{W}{R+(R-1)W}\right)(1-\kappa_C)$
	where $W = \frac{R^2\left(\overline{\mu^2}-(\overline{\mu})^2\right)}{(R-1)\mathrm{tr}(\Sigma)} \geq 0$	where $W = \frac{R^2\left(\overline{\mu^2}-(\overline{\mu})^2\right)}{(R-1)\mathrm{tr}(\Sigma)} \geq 0$
	$\kappa_C \geq \kappa_F$	$\kappa_F \leq \kappa_C$
	$\kappa_C = \kappa_F$ if $\mu_1 = \mu_2 = \cdots = \mu_R$	$\kappa_F = \kappa_C$ if $\mu_1 = \mu_2 = \cdots = \mu_R$

Note. Noninterchangeable raters imply that the chance correction uses rater-specific category probabilities, whereas interchangeable raters imply rater-invariant category probabilities. Both types may be suitable when the same raters rated all items. However, the latter type is needed otherwise. We have proved all results for interchangeable raters in the right column. Furthermore, we have proved the expressions for noninterchangeable raters in the left column, where the associated connections follow either trivially from the right column or via similar steps as those in Appendices A and C. We report all results to achieve a symmetric reference table.

agreement and is needed when raters vary across items. In contrast, noninterchangeability implies that the chance agreement uses *rater-specific* category probabilities, suitable when the same raters rated all items, and especially appropriate for intentionally assigned raters. However, less consensus may exist when the same but randomly selected raters (drawn from a larger population they are supposed to represent) classify all items, as the raters' identities are known but without being of interest. It is noteworthy that the most common coefficient for two raters (i.e., Cohen's kappa) assumes noninterchangeable raters, while the most common multirater coefficient (i.e., Fleiss' kappa) assumes interchangeable raters. Unless multirater settings differ substantially from two-rater settings (e.g., in terms of the same versus different raters across items), such a situation does not seem conceptually justified.

A crucial finding for quadratic weights is that the conceptual distinction between interchangeable and noninterchangeable raters becomes empirically irrelevant if the raters arrive at (approximately) the same mean rating, which one can assess quickly. Raters do *not* need (almost) identical distributions of their category proportions to become empirically interchangeable; being similar in only their mean ratings is sufficient. Especially in multirater settings (where the data no longer fit in a two-dimensional contingency table and often remain unreported), we recommend reporting the rater means and variances as parsimonious summary statistics of the data. These statistics facilitate diagnostics, such as whether the conceptual distinction between interchangeable and noninterchangeable raters matters (implying different agreement coefficients) and whether the agreement coefficients may behave like (average) Pearson product–moment correlations that measure association (if the raters have approximately the same means and variances). Furthermore, the reported means and variances allow external researchers to compute the values of unreported coefficients (i) without access to the underlying data set and (ii) without having coded up any agreement coefficients. For example, given the reported coefficient value of either Conger's or Fleiss' kappa, external researchers (unsatisfied with that choice) can immediately compute the value of the other agreement coefficient.

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

Both conceptually and mathematically, Conger's and Fleiss' quadratically weighted kappas (and their two-rater versions: Cohen's kappa and Scott's pi) look quite different. However, we demonstrated that they are linear transformations of each other, conditional on the (squared) pairwise rater differences in mean ratings and corresponding variances. Furthermore, this linear connection shows that quadratically weighted agreement coefficients with interchangeable raters (represented by Scott's pi and Fleiss' kappa) imply lower coefficient values than those with noninterchangeable raters (represented by Cohen's and Conger's kappas).

Since noninterchangeable raters are applicable only when the same raters rated the items, all results involving noninterchangeable raters (including comparison with interchangeable raters) require such settings. In contrast, our interpretations of coefficients with interchangeable raters (e.g., Fleiss' kappa) are relevant in both situations involving the same raters and situations with different raters across items. Although our population-level result that Fleiss' kappa equals the Pearson correlation and regression slope for two randomly selected raters holds generally, the corresponding sample-level result of concatenation requires the same number of raters per item. Furthermore, our empirical illustrations considered only such complete data. Future research may consider inference in situations with incomplete data.

Funding statement. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare none.

References

Almehrizi, R. S., & Emam, M. (2023). Asymptotic standard errors of intraclass correlation coefficients for two-way model. *Communications in Statistics: Simulation and Computation*, *52*, 2073–2092. https://doi.org/10.1080/03610918.2021.1897624

Andres, A. M., & Hernandez, M. A. (2020). Hubert's multi-rater kappa revisited. *British Journal of Mathematical and Statistical Psychology*, *73*, 1–22. https://doi.org/10.1111/bmsp.12167

Andres, A. M., & Hernandez, M. A. (2025). Estimators of various kappa coefficients based on the unbiased estimator of the expected index of agreements. *Advances in Data Analysis and Classification*, *19*, 177–207. https://doi.org/10.1007/s11634-024-00581-x

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*, 3–23. https://doi.org/10.2307/3315487

Barnhart, H. X., Haber, M., & Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, *58*, 1020–1027. https://doi.org/10.1111/j.0006-341X.2002.01020.x

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communication through limited response questioning. *The Public Opinion Quarterly*, *18*, 303–308. https://doi.org/10.1086/266520

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*, 687–699. https://doi.org/10.1177/001316448104100307

Brenner, H., & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, *7*, 199–202. https://doi.org/10.1097/00001648-199603000-00016

Brusco, M. J., Stahl, S., & Steinley, D. (2008). An implicit enumeration method for an exact test of weighted kappa. *British Journal of Mathematical and Statistical Psychology*, *61*, 439–452. https://doi.org/10.1348/000711007X227058

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *5*, 423–429. https://doi.org/10.1016/0895-4356(93)90018-V

Carrasco, J. L. (2010). A generalized concordance correlation coefficient based on the variance components generalized linear mixed models for overdispersed count data. *Biometrics*, *66*, 897–904. https://doi.org/10.1111/j.1541-0420.2009.01335.x

Carrasco, J. L., & Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, *59*, 849–858. https://doi.org/10.1111/j.0006-341X.2003.00099.x

Cicchetti, D. V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, *11*, 101–109. https://doi.org/10.1080/00029238.1971.11080840

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220. https://doi.org/10.1037/h0026256

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*, 322–328. https://doi.org/10.1037/0033-2909.88.2.322

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549. https://doi.org/10.1016/0895-4356(90)90158-l

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382. https://doi.org/10.1037/h0031619

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619. https://doi.org/10.1177/001316447303300309

Graham, P., & Jackson, R. (1993). The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology*, *46*, 1055–1062. https://doi.org/10.1016/0895-4356(93)90173-x

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29–48. https://doi.org/10.1348/000711006X126600

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. (4th ed.) Advanced Analytics, LLC.

Hoehler, F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, *53*, 499–503. https://doi.org/10.1016/s0895-4356(99)00174-2

Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement*, *24*, 749–753. https://doi.org/10.1177/001316446402400402

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, *84*, 289–297. https://doi.org/10.1037/0033-2909.84.2.289

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, *61*, 277–289. https://doi.org/10.1177/00131640121971239

Janson, H., & Olsson, U. (2004). A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement*, *64*, 62–70. https://doi.org/10.1177/0013164403260195

Janson, S., & Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research*, *14*, 255–269. https://doi.org/10.1207/s15327906mbr1402_9

King, T. S., & Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*, *20*, 2131–2147. https://doi.org/10.1002/sim.845

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*, 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. (4th ed.) Sage.

Kvålseth, T. O. (2018). An alternative interpretation of the linearly weighted kappa coefficients for ordinal data. *Psychometrika*, *83*, 618–627. https://doi.org/10.1007/s11336-018-9621-1

Landis, J. R., & Koch, G. G. (1977a). A one-way components of variance model for categorial data. *Biometrics*, *33*, 671–679. https://doi.org/10.2307/2529465

Landis, J. R., & Koch, G. G. (1977b). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. https://doi.org/10.2307/2529310

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*, 365–377. https://doi.org/10.1037/h0031643

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255–268. https://doi.org/10.2307/2532051

Lin, L. I.-K. (2000). A note on the concordance correlation coefficient. *Biometrics*, *56*, 324–325. https://doi.org/10.1111/j.0006-341X.2000.00324.x

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry*, *130*, 79–83. https://doi.org/10.1192/bjp.130.1.79

Mielke, P. W., Berry, K. J., & Johnston, J. E. (2007). The exact variance of weighted kappa with multiple raters. *Psychological Reports*, *101*, 655–660. https://doi.org/10.2466/pr0.101.2.655-660

Mielke, P. W., Berry, K. J., & Johnston, J. E. (2009). Unweighted and weighted kappa as measures of agreement for multiple judges. *International Journal of Management*, *26*, 213–223.

Moss, J. (2023). Measuring agreement using guessing models and knowledge coefficients. *Psychometrika*, *88*, 1002–1025. https://doi.org/10.1007/s11336-023-09919-4

Schouten, H. J. A. (1980). Measuring pairwise agreement among many observers. *Biometrical Journal*, *22*, 497–504. https://doi.org/10.1002/bimj.4710220605

Schouten, H. J. A. (1982). Measuring pairwise agreement among many observers. II. Some improvements and additions. *Biometrical Journal*, *24*, 431–435. https://doi.org/10.1002/bimj.4710240502

Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, *64*, 243–253. https://doi.org/10.1177/0013164403260197

Schuster, C., & Smith, D. A. (2005). Dispersion-weighted kappa: An integrative framework for metric and nominal scale agreement coefficients. *Psychometrika*, *70*, 135–146. https://doi.org/10.1007/s11336-003-1110-4

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321–325. https://doi.org/10.1086/266577
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420
Stoyan, D., Pommerening, A., Hummel, M., & Kopp-Schneider, A. (2018). Multiple-rater kappas for binary data: Models and interpretations. *Biometrical Journal*, *60*, 381–394. https://doi.org/10.1002/bimj.201600267
Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, *58*, 655–661. https://doi.org/10.1016/j.jclinepi.2004.02.021
Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa – A comparison of basic properties. *MethodsX*, *10*, 1–7. https://doi.org/10.1016/j.mex.2023.102212
Van Oest, R. (2019). A new coefficient of interrater agreement: The challenge of highly unequal category proportions. *Psychological Methods*, *24*, 439–451. https://doi.org/10.1037/met0000183
Van Oest, R. (2023). The dependence of chance-corrected weighted agreement coefficients on the power parameter of the weighting scheme: Analysis and measurement. *Psychometrika*, *88*, 554–579. https://doi.org/10.1007/s11336-022-09881-7
Van Oest, R., & Girard, J. M. (2022). Weighting schemes and incomplete data: A generalized Bayesian framework for chance-corrected interrater agreement. *Psychological Methods*, *27*, 1069–1088. https://doi.org/10.1037/met0000412
Vanbelle, S. (2016). A new interpretation of the weighted kappa coefficients. *Psychometrika*, *81*, 399–410. https://doi.org/10.1007/s11336-014-9439-4
Vanbelle, S. (2019). Asymptotic variability of (multilevel) multirater kappa coefficients. *Statistical Methods in Medical Research*, *28*, 3012–3026. https://doi.org/10.1177/0962280218794733
Vanbelle, S., & Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology*, *6*, 157–163. https://doi.org/10.1016/j.stamet.2008.06.001
Warrens, M. J. (2010a). Inequalities between kappa and kappa-like statistics for *kxk* tables. *Psychometrika*, *75*, 176–185. https://doi.org/10.1007/s11336-009-9138-8
Warrens, M. J. (2010b). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, *4*, 271–286. https://doi.org/10.1007/s11634-010-0073-4
Warrens, M. J. (2011). Cohen's linearly weighted kappa is a weighted average of 2x2 kappas. *Psychometrika*, *76*, 471–486. https://doi.org/10.1007/s11336-011-9210-z
Warrens, M. J. (2012a). Equivalences of weighted kappas for multiple raters. *Statistical Methodology*, *9*, 407–422. https://doi.org/10.1016/j.stamet.2011.11.001
Warrens, M. J. (2012b). Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. *Statistical Methodology*, *9*, 440–444. https://doi.org/10.1016/j.stamet.2011.08.006
Warrens, M. J. (2012c). Cohen's linearly weighted kappa is a weighted average. *Advances in Data Analysis and Classification*, *6*, 67–79. https://doi.org/10.1007/s11634-011-0094-7
Warrens, M. J. (2012d). Some paradoxical results for the quadratically weighted kappa. *Psychometrika*, *77*, 315–323. https://doi.org/10.1007/s11336-012-9258-4
Warrens, M. J. (2013). Conditional inequalities between Cohen's kappa and weighted kappas. *Statistical Methodology*, *10*, 14–22. https://doi.org/10.1016/j.stamet.2012.05.004
Warrens, M. J. (2014). Corrected Zegers-ten Berge coefficients are special cases of Cohen's weighted kappa. *Journal of Classification*, *31*, 179–193. https://doi.org/10.1007/s00357-014-9156-9
Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach*. (5th ed.) Cengage Learning
Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*, 374–378. https://doi.org/10.1037/0033-2909.103.3.374

Appendix A

We write Scott's quadratically weighted pi in (7) as a linear function of Cohen's quadratically weighted kappa in (3):

$$\kappa = \frac{2\sigma_{1,2}}{{\sigma_1}^2 + {\sigma_2}^2 + (\mu_1 - \mu_2)^2}.$$

This derivation is straightforward:

$$\begin{aligned}
\pi &= \frac{2\sigma_{1,2} - \frac{1}{2}(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2 + \frac{1}{2}(\mu_1 - \mu_2)^2} = \frac{2\sigma_{1,2}}{{\sigma_1}^2 + {\sigma_2}^2 + \frac{1}{2}(\mu_1 - \mu_2)^2} - \frac{\frac{1}{2}(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2 + \frac{1}{2}(\mu_1 - \mu_2)^2} \\
&= \kappa \times \frac{{\sigma_1}^2 + {\sigma_2}^2 + (\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2 + \frac{1}{2}(\mu_1 - \mu_2)^2} - \frac{\frac{1}{2}(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2 + \frac{1}{2}(\mu_1 - \mu_2)^2} \\
&= \kappa \times \frac{2\left({\sigma_1}^2 + {\sigma_2}^2\right) + 2(\mu_1 - \mu_2)^2}{2\left({\sigma_1}^2 + {\sigma_2}^2\right) + (\mu_1 - \mu_2)^2} - \frac{(\mu_1 - \mu_2)^2}{2\left({\sigma_1}^2 + {\sigma_2}^2\right) + (\mu_1 - \mu_2)^2}.
\end{aligned}$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

Dividing all numerators and denominators by ${\sigma_1}^2 + {\sigma_2}^2$ and substituting

$$W = \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2} \geq 0$$

yields the desired result:

$$\pi = \kappa \times \frac{2 + 2W}{2 + W} - \frac{W}{2 + W} = \kappa - \left(\frac{W}{2 + W}\right)(1 - \kappa).$$

Appendix B

We show that Fleiss' quadratically weighted kappa in (21) equals Lin's generalized concordance correlation coefficient after recentering the rater means and covariance matrix at the raters' grand mean. This kappa coefficient is

$$\kappa_F = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}} - \frac{1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r - \mu_{\tilde{r}})^2}{(R-1)\sum_{r=1}^{R}{\sigma_r}^2 + \frac{R-1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r - \mu_{\tilde{r}})^2}.$$

We can write the numerator in κ_F as

$$\begin{aligned}
&2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}} - \frac{1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r - \mu_{\tilde{r}})^2 \\
&= 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}} - \frac{1}{R}\left((R-1)\sum_{r=1}^{R}{\mu_r}^2 - 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\mu_r\mu_{\tilde{r}}\right) \\
&= 2\left(\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}} - \frac{R-1}{2R}\sum_{r=1}^{R}{\mu_r}^2 + \frac{1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\mu_r\mu_{\tilde{r}}\right) \\
&= 2\left(\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\sigma_{r,\tilde{r}} + \mu_r\mu_{\tilde{r}}) - \frac{R-1}{2R}\sum_{r=1}^{R}{\mu_r}^2 - \frac{R-1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\mu_r\mu_{\tilde{r}}\right) \\
&= 2\left(\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\sigma_{r,\tilde{r}} + \mu_r\mu_{\tilde{r}}) - \frac{R-1}{2R}\left(\sum_{r=1}^{R}{\mu_r}^2 + 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\mu_r\mu_{\tilde{r}}\right)\right) \\
&= 2\left(\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\sigma_{r,\tilde{r}} + \mu_r\mu_{\tilde{r}}) - \frac{R-1}{2R}\left(\sum_{r=1}^{R}\mu_r\right)^2\right) \\
&= 2\left(\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\sigma_{r,\tilde{r}} + \mu_r\mu_{\tilde{r}}) - \frac{R(R-1)}{2}\left(\frac{\sum_{r=1}^{R}\mu_r}{R}\right)^2\right) \\
&= 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\sigma_{r,\tilde{r}} + \mu_r\mu_{\tilde{r}} - \overline{\mu}^2\right),
\end{aligned}$$

where, in the last step, we used that $\overline{\mu} = \sum_{r=1}^{R}\mu_r/R$, and $R(R-1)/2$ is the number of elements in the double summation.

Similarly, we can write the denominator in κ_F as

$$\begin{aligned}
&(R-1)\sum_{r=1}^{R}{\sigma_r}^2 + \frac{R-1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r - \mu_{\tilde{r}})^2 \\
&= (R-1)\sum_{r=1}^{R}{\sigma_r}^2 + \frac{R-1}{R}\left((R-1)\sum_{r=1}^{R}{\mu_r}^2 - 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\mu_r\mu_{\tilde{r}}\right) \\
&= (R-1)\left(\sum_{r=1}^{R}{\sigma_r}^2 + \frac{R-1}{R}\sum_{r=1}^{R}{\mu_r}^2 - \frac{2}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\mu_r\mu_{\tilde{r}}\right) \\
&= (R-1)\left(\sum_{r=1}^{R}\left({\sigma_r}^2 + {\mu_r}^2\right) - \frac{1}{R}\sum_{r=1}^{R}{\mu_r}^2 - \frac{2}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\mu_r\mu_{\tilde{r}}\right) \\
&= (R-1)\left(\sum_{r=1}^{R}\left({\sigma_r}^2 + {\mu_r}^2\right) - \frac{1}{R}\left(\sum_{r=1}^{R}{\mu_r}^2 + 2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\mu_r\mu_{\tilde{r}}\right)\right) \\
&= (R-1)\left(\sum_{r=1}^{R}\left({\sigma_r}^2 + {\mu_r}^2\right) - \frac{1}{R}\left(\sum_{r=1}^{R}\mu_r\right)^2\right)
\end{aligned}$$

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press

$$= (R-1)\left(\sum_{r=1}^{R}\left(\sigma_r{}^2+\mu_r{}^2\right)-R\left(\frac{\sum_{r=1}^{R}\mu_r}{R}\right)^2\right)$$
$$= (R-1)\sum_{r=1}^{R}\left(\sigma_r{}^2+\mu_r{}^2-\overline{\mu}^2\right).$$

Substituting these new expressions for the numerator and denominator into κ_F yields

$$\kappa_F = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\left(\sigma_{r,\tilde{r}}+\mu_r\mu_{\tilde{r}}-\overline{\mu}^2\right)}{(R-1)\sum_{r=1}^{R-1}\left(\sigma_r{}^2+\mu_r{}^2-\overline{\mu}^2\right)} = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\tilde{\sigma}_{r,\tilde{r}}}{(R-1)\sum_{r=1}^{R}\tilde{\sigma}_r^2},$$

where $\tilde{\sigma}_{r,\tilde{r}} = \sigma_{r,\tilde{r}}+\mu_r\mu_{\tilde{r}}-\overline{\mu}^2$ and $\tilde{\sigma}_r^2 = \sigma_r{}^2+\mu_r{}^2-\overline{\mu}^2$ are the recentered covariances and variances, respectively.

Appendix C

We write Fleiss' quadratically weighted kappa in (21) as a linear function of Conger's quadratically weighted kappa in (15):

$$\kappa_C = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}}}{(R-1)\sum_{r=1}^{R}\sigma_r{}^2+\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r-\mu_{\tilde{r}})^2}.$$

We proceed as follows:

$$\kappa_F = \frac{2\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}\sigma_{r,\tilde{r}}-\frac{1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r-\mu_{\tilde{r}})^2}{(R-1)\sum_{r=1}^{R}\sigma_r{}^2+\frac{R-1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r-\mu_{\tilde{r}})^2}$$
$$= \kappa_C \times \frac{(R-1)\sum_{r=1}^{R}\sigma_r{}^2+\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r-\mu_{\tilde{r}})^2}{(R-1)\sum_{r=1}^{R}\sigma_r{}^2+\frac{R-1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r-\mu_{\tilde{r}})^2} - \frac{\frac{1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r-\mu_{\tilde{r}})^2}{(R-1)\sum_{r=1}^{R}\sigma_r{}^2+\frac{R-1}{R}\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r-\mu_{\tilde{r}})^2}.$$

Dividing all numerators and denominators by $(R-1)\sum_{r=1}^{R}\sigma_r{}^2$ and substituting

$$W = \frac{\sum_{r=1}^{R-1}\sum_{\tilde{r}>r}(\mu_r-\mu_{\tilde{r}})^2}{(R-1)\sum_{r=1}^{R}\sigma_r{}^2} \geq 0$$

yields the desired result:

$$\kappa_F = \kappa_C \times \frac{1+W}{1+\frac{R-1}{R}W} - \frac{\frac{1}{R}W}{1+\frac{R-1}{R}W} = \kappa_C \times \frac{R(1+W)}{R+(R-1)W} - \frac{W}{R+(R-1)W},$$

or equivalently,

$$\kappa_F = \kappa_C - \left(\frac{W}{R+(R-1)W}\right)(1-\kappa_C).$$

Cite this article: van Oest, R. and Moss, J. (2026). Quadratically Weighted Agreement Coefficients: Interpretations and Connections. *Psychometrika*, 1069–1091. https://doi.org/10.1017/psy.2026.10085

https://doi.org/10.1017/psy.2026.10085 Published online by Cambridge University Press