A COMPARISON OF COHEN'S KAPPA AND AGREEMENT COEFFICIENTS BY CORRADO GINI

Matthijs J. Warrens
Leiden University, Institute of Psychology, Unit Methodology and Statistics
P.O. Box 9555, 2300 RB Leiden, Email: warrens@fsw.leidenuniv.nl

ABSTRACT

The paper compares four coefficients that can be used to summarize inter-rater agreement on a nominal scale. The coefficients are Cohen's kappa and three coefficients that were originally proposed by the Italian statistician Corrado Gini. All four coefficients have zero value if the two nominal variables are statistically independent, and value unity if there is perfect agreement. The coefficients are compared both analytically and empirically. An ordering between the four coefficients is formally proved. It turns out that Cohen's kappa is a lower bound of the other coefficients. Moreover, it is shown that the point estimates of Cohen's kappa and the two smallest of Gini's coefficients are very similar for real data. We conclude that these three coefficients lead to the same conclusions about the degree of inter-rater agreement in practice.

Keywords: *Cohen's kappa; Inter-rater agreement; Nominal categories; Reliability coefficient; Corrado Gini.*

1. INTRODUCTION

In behavioral, health and engineering sciences it is frequently required that an observer assigns a group of objects (individuals) to a set of nominal (unordered, mutually exclusive) categories. The observer or rater may be a psychologist that classifies subjects on personality type, a clinician that classifies subjects on mental disorders, or an expert that classifies production faults [14,15,31,56]. Since there is often no golden standard researchers usually require that the rating task is performed by at least two raters. The agreement between the ratings of the two observers can then be used as an indicator of the quality of the category definitions and the raters' ability to apply them. Instead of studying and understanding the observed patterns of agreement and disagreement, researchers are often only interested in a single number that summarizes the degree of agreement. Various coefficients have been proposed that can be used to summarize the agreement between two raters on a nominal scale [42,47,56]. The most widely used coefficient is Cohen's kappa [5,9,22,45,46]. The popularity of kappa has led to the development of many extensions, including, kappas for three or more raters [11,48], kappas for groups of raters [38,39] and kappas for ordinal categories [49,50,51,52,53,54].

Cohen's kappa was originally proposed on an ad hoc basis as a descriptive statistic indicating degree of beyond-chance agreement [8,34,56]. Kraemer [26] showed that Cohen's kappa for two categories satisfies the classical definition of reliability. Although proposed as the proportion of agreement beyond chance [9], the value of kappa for three or more categories is generally considered to be uninterpretable, because no single coefficient is sufficient to completely and accurately convey information on agreement when there are more than two categories [27]. Furthermore, a general problem with agreement coefficients and other association coefficients is that often only the extreme values (maximum and zero values) have a clear interpretation [31].

Despite the difficulties with its interpretation, Cohen's kappa continues to be the most popular coefficient for summarizing inter-rater agreement on a nominal scale [22,56]. A main reason for kappa's popularity appears to be that its extreme values have a clear interpretation. Kappa has zero value when the two nominal variables (raters) are statistically independent and value unity if there is perfect agreement [9]. However, these properties are not unique to Cohen's kappa. Indeed, several authors have proposed agreement coefficients that have identical properties and it is therefore a moot point which coefficient is the best indicator of agreement of the ratings given these criteria.

In this paper we compare Cohen's kappa to three other agreement coefficients that have been proposed in the literature. It turns out that all three agreement coefficients were originally introduced by the Italian statistician Corrado Gini [16,17]. Gini's coefficients have been rediscovered by other authors [8,9,23,34]. The agreement coefficients are compared both analytically and empirically, and it is investigated whether the coefficients may lead to different conclusions in practice. The paper is organized as follows. Cohen's kappa is introduced in the next section. The three agreement coefficients originally proposed by Gini [16,17] are introduced in Section 3. An ordering between the four coefficients that is frequently observed in practice is formally proved in Section 4. Section 5 contains a discussion.

2. COHEN'S KAPPA

The literature contains a vast amount of coefficients for summarizing association or agreement between two nominal scale variables [12,19,35,40]. This paper is limited to coefficients for two nominal variables with *c* identical categories [8,23,46,47,56].

Suppose that two raters each independently classify the same set of objects (individuals, observations) into the same set of *c* categories that are defined in advance. For a population of *n* objects, let π_{ij} for $i,j = 1,\ldots,c$ denote the proportion of objects classified into category *i* by the first rater and into category *j* by the second rater. The square table $\{\pi_{ij}\}$ is also called an agreement table. Row and column totals of $\{\pi_{ij}\}$ are denoted by

$$\pi_{i+} = \sum_{j=1}^{c} \pi_{ij} \quad \text{and} \quad \pi_{+i} = \sum_{j=1}^{c} \pi_{ji},$$

and will be called the marginal totals of $\{\pi_{ij}\}$. The cell probabilities π_{ii} on the main diagonal of $\{\pi_{ij}\}$ indicate how many objects were put in the same categories by both raters. The square contingency table $\{\pi_{ij}\}$ can be seen as a cross-classification of two nominal variables with identical categories. The agreement coefficients discussed in this paper can also be used for summarizing agreement if we have *n* observers of one type paired with *n* observers of a second type, and each of the 2*n* observers assigns an object to one of *c* categories.

In the remainder of the paper the symbol $\sum \pi_{ii}$ is used as short notation for $\sum_{i=1}^{c} \pi_{ii}$. Cohen's kappa is defined as

$$\kappa = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}} \tag{1}$$

where $\sum \pi_{ii}$ and $\sum \pi_{i+}\pi_{+i}$ are, respectively, the proportions of observed and expected agreement. The numerator $\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}$ is equal to zero if the ratings are statistically independent. Division by the denominator $1 - \sum \pi_{i+}\pi_{+i}$ sets the maximum value of kappa at unity.

Assuming a multinominal sampling model with the total numbers of objects *n* fixed, the maximum likelihood estimate of the cell probability π_{ij} is given by $\hat{\pi}_{ij} = n_{ij}/n$, where n_{ij} is the observed frequency. The maximum likelihood estimate $\hat{\kappa}$ of κ in (1) is obtained by replacing the cell probabilities π_{ij} by the $\hat{\pi}_{ij}$ [7]. An example of an observed agreement table $\{n_{ij}\}$ is presented in Table 1. The data in Table 1 are taken from Cohen [9]. In this study, 200 sets of fathers and mothers were asked to identify which of three personality descriptions (Types 1, 2 or 3) best describes their oldest child. Table 1 is the cross classification of the fathers description and mothers description of the oldest child. For the data in Table 1 we have

$$\sum \hat{\pi}_{ii} = 0.44 + 0.20 + 0.06 = 0.70,$$

and

$$\sum \hat{\pi}_{i+}\hat{\pi}_{+i} = (0.50)(0.60) + (0.30)^2 + (0.20)(0.10) = 0.41,$$

and the estimate $\hat{\kappa} = (0.70 - 0.41)/(1 - 0.41) = 0.492$, which indicates a moderate degree of agreement [28].

Table 1: Personality descriptions of oldest child by 200 sets of fathers and mothers [9].

	Mother			
Father	Type 1	Type 2	Type 3	Totals
Type 1	88	10	2	100
Type 2	14	40	6	60
Type 3	18	10	12	40
Totals	120	60	20	200

The maximum value of $\sum \pi_{ii}$ is restrained by the marginal totals in the sense that the value of π_{ii} cannot exceed the minimum of π_{i+} and π_{+i} [8,9,34]. For fixed marginal totals π_{i+} and π_{+i}, the maximum value of $\sum \pi_{ii}$ is given by

$$\max\left(\sum \pi_{ii}\right) = \sum \min\{\pi_{i+}, \pi_{+i}\}.$$

Replacing the 1 by $\max(\sum \pi_{ii})$ in definition (1) we obtain

$$\kappa_m = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{\sum \min\{\pi_{i+}, \pi_{+i}\} - \sum \pi_{i+}\pi_{+i}}. \tag{2}$$

This coefficient may be interpreted as kappa/max(kappa): κ_m is equal to Cohen's kappa divided by the maximum value of kappa given the marginal totals [8,9,13,34]. The value of κ_m is 1 when all objects that are assigned to category *i* by the first rater, are also assigned to category *i* by the second rater, or vice versa. Similar to Cohen's

kappa, the value of κ_m in (2) is zero when the two nominal variables are statistically independent. For the data in Table 1 we have the point estimate $\hat{\kappa}_m = 0.592$.
The special case of the coefficient κ_m for $c = 2$ categories [43,44] is discussed in Johnson [24] and Loevinger [29,30]. The latter author calls it coefficient *H*. Loevinger's *H* is a central coefficient in Mokken scale analysis, a methodology that can be used to select a subset of binary test items that are sensitive to the same underlying dimension [36]. Goodman and Kruskal [19,20] note that this special case was independently proposed in Benini [6] and Jordan [25]. Furthermore, in the case of positive agreement the special case of kappa/max(kappa) is equivalent to a coefficient discussed in Cole [10] and Zysno [57]. Moreover, for $c = 2$ categories kappa/max(kappa) is equivalent to phi/max(phi) [44]. A detailed review of the phi/max(phi) literature is presented in Davenport and El-Sanhurry [13].

3. AGREEMENT COEFFICIENTS BY GINI

From 1914 to 1916 the Italian statistician Corrado Gini published several papers in which he proposed a great variety of association coefficients. He examined in detail many distinctions between relationships within a bivariate distribution and proposed coefficients of association for the different cases, including several coefficients for agreement. Gini is best known for the Gini [18] coefficient, which is a coefficient of statistical dispersion that can be used as a coefficient of inequality of income or wealth. An exposition of the Gini material in English can be found in Weida [55]. The Gini material is also briefly reviewed in Goodman and Kruskal [19,20]. Goodman and Kruskal [19, p. 137] note that they, and we quote, ``have not found in Gini's papers operational interpretations of his proposed coefficients. They all seem to be of a formal nature in which consideration of absolute or quadratic differences, followed by averaging, is taken as reasonable without argument. Special attention is paid to denominators so as to make the indices range between 0 and 1 within appropriate limitations for variation in the joint distribution."
Goodman and Kruskal [19, p. 137] report that Gini [17] proposed the coefficient

$$G_1 = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i} - \frac{1}{2}\sum|\pi_{i+} - \pi_{+i}|}. \quad (3)$$

The numerator of G_1 in (3) is identical to the numerators of Cohen's kappa and kappa/max(kappa). The denominator of $G_1$is quite similar to that of kappa, although on first sight it is unclear why it is defined like this. However, the following theorem shows that $G_1$is in fact equivalent to kappa/max(kappa).

Theorem 1. $G_1 = \kappa_m$.
Proof: Since G_1 and κ_m have the same numerator it must be shown that the two denominators are equivalent. We have the identities

$$\frac{1}{2}\Big(\sum \max\{\pi_{i+}, \pi_{+i}\} + \sum \min\{\pi_{i+}, \pi_{+i}\}\Big) = \frac{1}{2}\Big(\sum \pi_{i+} + \sum \pi_{+i}\Big) = 1, \quad (4)$$

and

$$\frac{1}{2}\Big(\sum \max\{\pi_{i+}, \pi_{+i}\} - \sum \min\{\pi_{i+}, \pi_{+i}\}\Big) = \frac{1}{2}\sum|\pi_{i+} - \pi_{+i}|. \quad (5)$$

Subtracting (5) from (4) we obtain the identity

$$1 - \frac{1}{2}\sum|\pi_{i+} - \pi_{+i}| = \sum \min\{\pi_{i+}, \pi_{+i}\}.$$

Hence the denominators of G_1 and κ_m are equivalent. ■

Goodman and Kruskal [19, p. 137] report that Gini [16] proposed the coefficient

$$G_2 = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{\sqrt{(1 - \sum \pi_{i+}^2)(1 - \sum \pi_{+i}^2)}}. \quad (6)$$

Coefficient G_2 was independently proposed by Janson and Vegelius [23]. The statistic is a generalization of the phi coefficient for 2×2 tables [43,44] to the case of *c* nominal categories. Thus for $c = 2$ categories G_2 is similar to the Pearson correlation coefficient in its interpretation. Janson and Vegelius [23] do not provide an operational interpretation of G_2 for $c = 3$ categories. Similar to Cohen's kappa, the value of G_2 in (6) is unity when perfect agreement between the two raters occurs, and zero when agreement is equal to that expected under independence. For the data in Table 1 we have the point estimate $\hat{G}_2 = 0.501$.
Weida[55] reports that Gini also proposed the coefficient

$$G_3 = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \frac{1}{2}\sum \pi_{i+}^2 - \frac{1}{2}\sum \pi_{+i}^2}.$$

Coefficient G_3 was independently proposed by Popping [34, p.76]. The coefficient is a generalization of a coefficient by Maxwell and Pilliner [32] for the case of $c = 2$ categories. Popping [34] does not provide a physical meaning of G_3, but showed that both G_3 and κ satisfy a whole range of desirable properties. Again, the value of G_3 is unity when there is perfect agreement between the two raters, and zero when agreement is equal to that expected under independence. For the data in Table 1 we have the estimate $\hat{G}_3 = 0.500$.

4. INEQUALITIES

In the previous section we observed the ordering $\hat{G}_1 > \hat{G}_2 > \hat{G}_3 > \hat{\kappa} > 0$ for the data in Table 1. It turns out that this ordering of the values of the agreement coefficients is observed quite frequently in practice (see Table 2 in Section 5). Theorem 2 below shows that the triple inequality $|G_1| \geq |G_2| \geq |G_3| \geq |\kappa|$ holds, where the symbol $|A|$ denotes the absolute value of the coefficient A.

The inequality $|G_2| \geq |\kappa|$ is mentioned in Janson and Vegelius [23, p. 265] but no formal proof is provided. For $c = 2$ categories the inequality $|G_2| \geq |\kappa|$ was proved in Cohen [9] and Warrens [41]. For $c = 2$ categories the inequality $|G_3| \geq |\kappa|$ was proved in Warrens [41].

Theorem 2. $|G_1| \geq |G_2| \geq |G_3| \geq |\kappa|$.

Proof: We first prove the left inequality, then the middle, and finally the right inequality.

We have the identity

$$1 - \sum \pi_{i+}^2 = \sum \pi_{i+} - \sum \pi_{i+}^2 = \sum \pi_{i+}(1 - \pi_{i+}).$$

Hence, using the positive numbers

$$a_i = \sqrt{\pi_{i+}(1 - \pi_{i+})} \qquad \text{and} \qquad b_i = \sqrt{\pi_{+i}(1 - \pi_{+i})}$$

in the Cauchy-Schwarz inequality ([1, p. 11] or [33, p. 20])

$$\sum a_i^2 \sum b_i^2 \geq \left(\sum a_i b_i\right)^2,$$

yields

$$\sqrt{\left(1 - \sum \pi_{i+}^2\right)\left(1 - \sum \pi_{+i}^2\right)} \geq \sum \sqrt{\pi_{i+}(1 - \pi_{i+})\pi_{+i}(1 - \pi_{+i})}. \tag{7}$$

Furthermore, since the smallest of two real numbers never exceeds the geometric mean of the numbers, we have

$$\sqrt{\pi_{i+}(1 - \pi_{i+})\pi_{+i}(1 - \pi_{+i})} \geq \min\{\pi_{i+}(1 - \pi_{+i}), \pi_{+i}(1 - \pi_{i+})\}. \tag{8}$$

Summing (8) over all i, we obtain

$$\begin{aligned} \sum \sqrt{\pi_{i+}(1 - \pi_{i+})\pi_{+i}(1 - \pi_{+i})} &\geq \sum \min\{\pi_{i+}(1 - \pi_{+i}), \pi_{+i}(1 - \pi_{i+})\} \\ &= \sum \min\{\pi_{i+}, \pi_{+i}\} - \sum \pi_{i+}\pi_{+i}. \end{aligned} \tag{9}$$

Combining (7) and (9), we obtain

$$\sqrt{\left(1 - \sum \pi_{i+}^2\right)\left(1 - \sum \pi_{+i}^2\right)} \geq \sum \min\{\pi_{i+}, \pi_{+i}\} - \sum \pi_{i+}\pi_{+i}.$$

Hence, the denominator of G_1 never exceeds the denominator of G_2, and we conclude that $|G_1| \geq |G_2|$.

Next, inequality $|G_2| \geq |G_3|$ if and only if

$$\left(1 - \sum \pi_{i+}^2\right)\left(1 - \sum \pi_{+i}^2\right) \leq \left(1 - \frac{1}{2}\sum \pi_{i+}^2 - \frac{1}{2}\sum \pi_{+i}^2\right)^2. \tag{10}$$

Define

$$x = 1 - \sum \pi_{i+}^2 \quad \text{and} \quad y = 1 - \sum \pi_{+i}^2.$$

Then, inequality (10) can be written as

$$xy = \frac{(x+y)^2}{4} \qquad \text{or} \qquad 0 \leq \frac{(x-y)^2}{4}.$$

Hence, the denominator of G_2 never exceeds the denominator of G_3, and we conclude that $|G_2| \geq |G_3|$.

Finally, from the inequality $(\pi_{i+} - \pi_{+i})^2 \geq 0$,it follows that

$$\frac{\pi_{i+}^2 + \pi_{+i}^2}{2} \geq \pi_{i+}\pi_{+i}. \tag{11}$$

Summing (11) over all i, we obtain

$$\frac{1}{2}\sum \pi_{i+}^2 + \frac{1}{2}\sum \pi_{+i}^2 \geq \sum \pi_{i+}\pi_{+i}$$

which is equal to

$$1 - \frac{1}{2}\sum \pi_{i+}^2 - \frac{1}{2}\sum \pi_{+i}^2 \leq 1 - \sum \pi_{i+}\pi_{+i}.$$

Hence, the denominator of G_3 never exceeds the denominator of κ, and we conclude that $|G_3| \geq |\kappa|$. This completes the proof of the theorem. ■

Table 2: Point estimates of the coefficients G_1, G_2, G_3 and κ for 34 agreement tables from the literature. The values are sorted on the number of categories (# cat) and the value of $\hat{G}_1$.

Source of data	# cat	$\hat{G}_1$	$\hat{G}_2$	$\hat{G}_3$	$\hat{\kappa}$
[37, p. 307]	3	0.756	0.730	0.730	0.730
[9]	3	0.592	0.501	0.500	0.492
[2, p. 32]	3	0.445	0.369	0.369	0.363
[7, p. 397]	3	0.405	0.364	0.364	0.362
[7, p. 288]	3	0.320	0.244	0.244	0.236
[37, p. 301]	3	0.203	0.174	0.173	0.171
[3, p. 271]	3	0.243	0.158	0.158	0.140
[2, p. 360]	4	0.950	0.935	0.935	0.935
[2, p. 26]	4	0.814	0.738	0.737	0.735
[3, p. 270]	4	0.879	0.800	0.800	0.798
[4]	4	0.785	0.744	0.744	0.744
[3, p. 269]	4	0.761	0.674	0.674	0.668
[21, p. 170]	4	0.626	0.583	0.583	0.582
[37, p. 303]	4	0.649	0.552	0.552	0.545
[2, p. 376]	4	0.607	0.595	0.595	0.595
[3, p. 260]	4	0.792	0.536	0.532	0.493
[21, p. 170]	4	0.446	0.433	0.433	0.433
[4]	4	0.646	0.407	0.406	0.368
[2, p. 377]	4	0.480	0.392	0.392	0.385
[28]	4	0.408	0.308	0.308	0.297
[3, p. 272]	4	0.256	0.232	0.232	0.231
[21, p. 170]	4	0.332	0.225	0.224	0.208
[3, p. 273]	4	0.301	0.204	0.204	0.190
[2, p. 32]	4	0.147	0.131	0.131	0.129
[37, p. 303]	4	0.174	0.116	0.116	0.110
[3, p. 270]	4	0.289	0.117	0.109	0.077
[37, p. 272]	5	0.926	0.913	0.913	0.913
[7, p. 399]	5	0.901	0.861	0.861	0.860
[2, p. 368]	5	0.795	0.540	0.535	0.498
[7, p. 100]	5	0.229	0.216	0.216	0.215
[2, p. 376]	5	0.196	0.183	0.183	0.182
[37, p. 302]	6	0.564	0.540	0.540	0.539
[37, p. 277]	7	0.152	0.142	0.142	0.142
[37, p. 274]	8	0.608	0.418	0.417	0.393

5. DISCUSSION

In this paper we compared four coefficients of inter-rater agreement for nominal categories. The coefficients are Cohen's kappa and three coefficients (denoted by G_1, G_2 and G_3) that can be traced back to publications by the Italian statistician Corrado Gini. It was shown that G_1 is equivalent to kappa/max(kappa). All four agreement coefficients have zero value if the two nominal variables are statistically independent, and value unity if there is perfect agreement. However, for all four coefficients only the extreme values (maximum and zero values) have a clear interpretation. The meaning of the values between 1 and 0 is generally considered uninterpretable, because no single coefficient is sufficient to completely and accurately convey information on agreement when there are three or more categories [27]. Nevertheless, in practice most agreement studies are simply summarized by reporting a single value. The most popular agreement coefficient is Cohen's kappa [5,22,56]. Since there is no apparent reason for preferring kappa over Gini's coefficients, one may wonder whether the different coefficients also lead to different conclusions about the degree of agreement.

Table 2 presents the point estimates of G_1, G_2, G_3 and κ for 34 agreement tables from the literature. Close inspection reveals that the estimates of G_2 and G_3 are indistinguishable for most of the data entries. Moreover, the estimates of κ, G_2 and G_3 are very similar for all 34 data entries of Table 2. We conclude that the use of G_2, G_3 and κ will lead to the same conclusions about the degree of inter-rater agreement in practice. Only the use of coefficient $G_1 =$ kappa/max(kappa) may lead to other conclusions about the data. G_1 is commonly interpreted as the proportion of marginally permitted agreement beyond chance, whereas Cohen's κ, $G_2$and G_3 can be interpreted as the proportion of agreement beyond chance. For each 34 data entries of Table 2 we observe the ordering $G_1 > G_2 > G_3 > \kappa > 0$. The triple inequality $|G_1| \geq |G_2| \geq |G_3| \geq |\kappa|$ was proved in Section 4.

6. **ACKNOWLEDGEMENT**
This research is part of Veni project 451-11-026 funded by the Netherlands Organisation for Scientific Research.

REFERENCES

[1]. M. Abramowitz and I. A. Stegun. Handbook of Mathematical Functions (with Formulas, Graphs and Mathematical Tables). Dover Publications, New York, 1965.
[2]. A. Agresti. Categorical Data Analysis. Wiley, New York, 1990.
[3]. A. Agresti. An Introduction to Categorical Data Analysis. Wiley, New York, 2007.
[4]. M. Aickin. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. Biometrics, 46:293-302, 1990.
[5]. M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha. Beyond kappa: A review of interrater agreement measures. The Canadian Journal of Statistics, 27:3-23, 1999.
[6]. R. Benini. Principii di Demografia. G. Barbèra, Firenzi. No. 29 of Manuali Barbèra di Scienza Giuridiche Sociale e Politiche, 1901.
[7]. Y. M. Bishop, S. E. Fienberg, and P. W. Holland. Discrete Multivariate Analysis. Theory and Practice. MIT Press, Cambridge, 1975.
[8]. R. L. Brennan and D. J. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement, 41:687-699, 1981.
[9]. J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20:37-46, 1960.
[10]. L. C. Cole. The measurement of interspecific association. Ecology, 30:411-424, 1949.
[11]. A. J. Conger. Integration and generalization of kappas for multiple raters. Psychological Bulletin, 88:322-328, 1980.
[12]. H. Cramér. Mathematical Methods of Statistics. Princeton University Press, Princeton, 1946.
[13]. E. C. Davenport and N. A. El-Sanhurry. Phi/phimax: Review and synthesis. Educational and Psychological Measurement, 51:821-828, 1991.
[14]. J. L. Fleiss. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76:378-382, 1971.
[15]. J. L. Fleiss. Measuring agreement between two judges on the presence or absence of a trait. Biometrics, 31:651-659, 1975.
[16]. C. Gini. Indice di omofilia e di rassomiglianza e loro relazioni col coefficiente di correlazione e con gli indici di attrazione. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti, Series 8, 74:583-610, 1914-1915.
[17]. C. Gini. Nuovi contribute alla teoria delle relazioni statistiche. Attidel Reale Istituto Veneto di Scienze, Lettere ed Arti, Series 8, 74:1903-1942, 1914-1915.
[18]. C. Gini. Variabilità e mutabilità. In E. Pizetti and T. Salvemini, editors, Memorie di Metodologica Statistica. Libreria Eredi Virgilio Veschi, Rome, 1955.
[19]. L. A. Goodman and W. H. Kruskal. Measures of association for cross classifications II: Further discussion and references. Journal of the American Statistical Association, 54:123-163, 1959.
[20]. L. A. Goodman and W. H. Kruskal. Measures of Association for Cross Classifications. Springer-Verlag, New York, 1979.
[21]. D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski. A Handbook of Small Data Sets. Chapman & Hall/CRC, Boca Raton, 1994.
[22]. L. M. Hsu and R. Field. Interrater agreement measures: Comments on kappa_n, Cohen's kappa, Scott's π and Aickin's α. Understanding Statistics, 2:205-219, 2003.
[23]. S. Janson and J. Vegelius. On generalizations of the G index and the Phi coefficient to nominal scales. Multivariate Behavioral Research, 14:255-269, 1979.
[24]. H. M. Johnson. Maximal selectivity, correctivity and correlation obtainable in a 2×2 contingency table. American Journal of Psychology, 58:65-68, 1945.
[25]. K Jordan. A Korreláció számitása I. Magyar Statisztikai Szemle Kiadványai, 1. Szám, 1941.
[26]. H. C. Kraemer. Ramifications of a population model for κ as a coefficient of reliability. Psychometrika, 44:461-472, 1979.

[27]. H. C. Kraemer, V. S. Periyakoil, and A. Noda. Kappa coefficients in medical research. Statistics in Medicine, 21:2109-2129, 2002.
[28]. J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. Biometrics, 33:159-174, 1977.
[29]. J. Loevinger. The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. Psychological Bulletin, 45:507-529, 1948.
[30]. J. A. Loevinger. A Systematic Approach to the Construction and Evaluation of Tests of Ability. Psychological Monographs: General and Applied, Vol. 61, No. 4, 1947.
[31]. J. De Mast. Agreement and kappa-type indices. The American Statistician, 61:148-153, 2007.
[32]. A. E. Maxwell and A. E. G. Pilliner. Deriving coefficients of reliability and agreement for ratings. British Journal of Mathematical and Statistical Psychology, 21:105-116, 1968.
[33]. D. S. Mitrinović. Elementary Inequalities. P. Noordhoff Ltd., Groningen,1964.
[34]. R. Popping. Overeenstemmingsmaten voor Nominale Data. Rijksuniversiteit Groningen, Groningen, 1983.
[35]. C. E. Särndal. A comparative study of association measures. Psychometrika, 39:165-187, 1974.
[36]. K. Sijtsma and I. W. Molenaar. Introduction to Nonparametric Item Response Theory. Sage, Thousand Oaks, 2002.
[37]. J. S. Simonoff. Analyzing Categorical Data. Springer-Verlag, New York,2003.
[38]. S. Vanbelle and A. Albert. Agreement between an isolated rater and a group of raters. Statistica Neerlandica, 63:82-100, 2009.
[39]. S. Vanbelle and A. Albert. Agreement between two independent groups of raters. Psychometrika, 74:477-491, 2009.
[40]. J. Vegelius and S. Janson. Criteria for symmetric measures of association for nominal data. Quality and Quantity, 16:243-250, 1982.
[41]. M. J. Warrens. Bounds of resemblance measures for binary (presence/absence) variables. Journal of Classification, 25:195-208, 2008.
[42]. M. J. Warrens. On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. Psychometrika, 73:777-789, 2008.
[43]. M. J. Warrens. On similarity coefficients for 2×2 tables and correction for chance. Psychometrika, 73:487-502, 2008.
[44]. M. J. Warrens. On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. Journal of Classification, 25:177-183, 2008.
[45]. M. J. Warrens. Cohen's kappa can always be increased and decreased by combining categories. Statistical Methodology, 7:673-677, 2010.
[46]. M. J. Warrens. A formal proof of a paradox associated with Cohen's kappa. Journal of Classification, 27:322-332, 2010.
[47]. M. J. Warrens. Inequalities between kappa and kappa-like statistics for $k\times k$ tables. Psychometrika, 75:176-185, 2010.
[48]. M. J. Warrens. Inequalities between multi-rater kappas. Advances in Data Analysis and Classification, 4:271-286, 2010.
[49]. M. J. Warrens. Cohen's kappa is a weighted average. Statistical Methodology, 8:473-484, 2011.
[50]. M. J. Warrens. Cohen's linearly weighted kappa is a weighted average of 2×2 kappas. Psychometrika, 76:471-486, 2011.
[51]. M. J. Warrens. Cohen's linearly weighted kappa is a weighted average. Advances in Data Analysis and Classification, 6:67-79, 2012.
[52]. M. J. Warrens. Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. Statistical Methodology, 9:440-444, 2012.
[53]. M. J. Warrens. Cohen's weighted kappa with additive weights. Advances in Data Analysis and Classification, 7:41-55, 2013.
[54]. M. J. Warrens. Conditional inequalities between Cohen's kappa and weighted kappas. Statistical Methodology, 10:14-22, 2013.
[55]. F. M. Weida. On various concepts of correlation. Annals of Mathematics, 29:276-312, 1928.
[56]. R. Zwick. Another look at interrater agreement. Psychological Bulletin, 103:374-378, 1988.
[57]. P. V. Zysno. The modification of the Phi-coefficient reducing its dependence on the marginal distributions. Methods of Psychological Research Online, 2:41-52, 1997.