AGREEMENT BETWEEN TWO INDEPENDENT GROUPS OF RATERS

February 20, 2009

Correspondence should be sent to

Sophie Vanbelle
Department of Biostatistics
School of Public Health
University of Lige, CHU Sart Tilman (B23),
4000 Lige, Belgium
+32 4 366 2590
sophie.vanbelle@ulg.ac.be

Abstract

We propose a coefficient of agreement to assess the degree of concordance between two independent groups of raters classifying items on a nominal scale. This coefficient, defined on a population-based model, extends the classical Cohen's kappa coefficient for quantifying agreement between two raters. Weighted and intraclass versions of the coefficient are also given and their sampling variance is determined by the Jackknife method. The method is illustrated on medical education data which motivated the research.

Key words: Agreement, nominal scale, kappa coefficient

1. Introduction

Kappa-like agreement indexes are commonly used to quantify agreement between two raters on a categorical scale. They include Cohen's kappa coefficient (Cohen, 1960), the weighted kappa coefficient (Cohen, 1968) and the intraclass kappa coefficient (Kraemer, 1979) which was extended to several raters by Fleiss (1981). All coefficients are based on the same principle: the proportion of concordant classifications between the two raters (p_o) is corrected for the proportion of concordant classifications expected by chance (p_e) and standardized $\hat{\kappa} = (p_o - p_e)/(1 - p_e)$ to obtain a value 1 when agreement between the two raters is perfect and 0 in case of agreement due to chance alone. Although agreement is often searched between two individual raters, there are situations where agreement is needed between two groups of raters. For example, a group of students may be evaluated against another group of students or against a group of experts, each group classifying the same set of items on a categorical scale. Likewise, agreement may be searched between two groups of physicians with different specialties or professional experience in diagnosing patients by means of the same (positive/negative) clinical test. In such instances, each group is seen as a whole, a global entity with its own heterogeneity. Interest resides in the overall degree of agreement between the groups, not in the agreement between individuals themselves. In fact, the groups may perfectly agree while some of their members may not.

Methods testing for evidence of agreement between two groups of raters when ordering items were proposed by Schucany & Frawley (1973), Hollander & Sethuraman (1978), Kraemer (1981) and Feigin & Alvo (1986). These methods are generally based on the Spearman rank correlation coefficient or Kendall's tau coefficient. However, methods designed to quantify the degree of agreement between two groups of raters on a nominal or ordinal scale barely exist and it appears that the only reference found in the literature is a paper written by Schouten (1982). He developed a measure of pairwise interobserver agreement between two groups of raters to find clusters of homogeneous subgroups of raters when all raters classify the items on a categorical scale. His method consists in substituting in the kappa coefficient the observed proportion of agreement (p_o) and the proportion of agreement expected by chance (p_e) by, respectively, the mean of the observed ($\overline{p}_o$) and of the expected ($\overline{p}_e$) proportions of agreement obtained between all possible pairs of raters formed with one rater in each group, namely $\kappa = (\overline{p}_o - \overline{p}_e)/(1 - \overline{p}_e)$. Unfortunately, in Schouten's approach, perfect agreement between the two groups can only be achieved if there is perfect agreement within each group . Although there is a clear lack of theoretical work on agreement measures between two groups of raters, it is common practice in the applied literature to determine empirically a consensus category in each group of raters in order to reduce the problem to the case of two raters. To our knowledge, the consensus method is used as an intuitive method and there is no theoretical proof to justify its use. The consensus category may be defined as the modal category (e.g., van Hoeij & al., 2004), the median category (e.g., Raine & al., 2004) or the mean category (e.g., Bland & al., 2005) if the scale is ordinal. When a consensus category is found in each group for each item, the agreement between these categories is studied in the usual way (case of two raters). In all instances, however, the question of how to proceed when a consensus can not be reached remains. Moreover, different rules to define the consensus category may lead to different conclusions (Kraemer & al., 2004). Indeed, consider a group of 10 raters allocating an item on a 5-point Likert scale and suppose that 3 raters classify the item in category 1, 2 in category 2, none in categories 3 and 4, and 5 in category 5. The consensus category defined by the modal rule is category 5, by the median rule category 2, 3, 4 or 5 and by the mean rule category 3 (category chosen by none of the raters in the group). The three rules may almost inevitably lead to three different conclusions. It should also be remarked that consensus does not take into account the variability in the groups in the sense that different patterns of responses may lead to the determination of the same consensus category and thus lead to the same conclusions. Indeed, in the example above, if 6 instead of 5 raters classified the item in category 5, the modal category would still be category 5, leading to the same conclusion although the variability in the group is different.

The present research study aimed at defining an overall agreement index between two groups of raters, taking into account the heterogeneity of each group. Furthermore, the agreement index overcomes the problem of consensus and can be viewed as a natural extension of Cohen's kappa coefficient to two groups of raters. The novel agreement index is defined on a population-based model and its sampling variability is determined by the Jackknife method (Efron & Tibshirani, 1993).

2. Agreement within populations of raters

Consider a population $\mathcal{I}$ of items and two populations $\mathcal{R}_g$ of raters $(g = 1, 2)$. Suppose that items have to be classified in two categories $(K = 2)$. Now, consider a randomly selected rater r from population $\mathcal{R}_g$ and a randomly selected item i from population $\mathcal{I}$. Let $X_{ir,g}$ be the random variable such that $X_{ir,g} = 1$ if rater r of population $\mathcal{R}_g$ classifies item i in category 1 and $X_{ir,g} = 0$ otherwise. For each item i, $E(X_{ir,g}|i) = P(X_{ir,g} = 1) = P_{i,g}$ over the population of raters. Then, over the population of items, $E(P_{i,g}) = E[E(X_{ir,g}|i)] = \pi_g$ and $var(P_{i,g}) = \sigma_g^2$. The agreement within the population of raters $\mathcal{R}_g$ is classically quantified by the intraclass coefficient (Kraemer & al., 1979), denoted ICC_g,

$$ICC_g = \frac{\sigma_g^2}{\pi_g(1 - \pi_g)} \tag{1}$$

It is easily shown that $0 \leq ICC_g \leq 1$. The value $ICC_g = 1$ corresponds to perfect agreement within the population of raters. By contrast, $ICC_g = 0$ when heterogeneity of the items is not detected by the raters or when items are homogeneous in the population (Kraemer & al., 2004).

3. Definition of the agreement index

3.1. Dichotomous scale (K=2)

Using the notation above, we suppose that the two populations of raters $\mathcal{R}_1$ and $\mathcal{R}_2$ have to independently classify a randomly chosen item i from population $\mathcal{I}$ in two categories $(K = 2)$. The joint distribution of the classifications of item i made by the two populations of raters consists of four probabilities summing up to 1, $(1 - P_{i,1})(1 - P_{i,2})$, $(1 - P_{i,1})P_{i,2}$, $P_{i,1}(1 - P_{i,2})$ and $P_{i,1}P_{i,2}$. For example, $P_{i,1}P_{i,2}$ denotes the probability that both populations of raters classify item i into category 1. The expectations of these joint probabilities over the population of items $\mathcal{I}$ can be represented in a 2×2 classification table, as displayed in Table 1 with $\rho = corr(P_{i,1}, P_{i,2}) = [E(P_{i,1}P_{i,2}) - \pi_1\pi_2)]/\sigma_1\sigma_2$, the correlation over $\mathcal{I}$ between the random variables $P_{i,1}$ and $P_{i,2}$.

The probability that the two populations of raters agree on the classification of item i is naturally defined by

$$\Pi_i = P_{i,1}P_{i,2} + (1 - P_{i,1})(1 - P_{i,2}) \tag{2}$$

Table 1.

Expected joint classification probabilities of the two populations of raters over the population of items

		$\mathcal{R}_2$		
		0	1	
$\mathcal{R}_1$	0	$E[(1-P_{i,1})(1-P_{i,2})]$ $(1-\pi_1)(1-\pi_2)+\rho\sigma_1\sigma_2$	$E[(1-P_{i,1})P_{i,2}]$ $(1-\pi_1)\pi_2-\rho\sigma_1\sigma_2$	$1-\pi_1$
	1	$E[P_{i,1}(1-P_{i,2})]$ $\pi_1(1-\pi_2)-\rho\sigma_1\sigma_2$	$E[P_{i,1}P_{i,2}]$ $\pi_1\pi_2+\rho\sigma_1\sigma_2$	π_1
		$1-\pi_2$	π_2	1

Thus, at the population level, the *mean probability of agreement* over $\mathcal{I}$ is (see Table 1)

$$\Pi_T = E(\Pi_i) = \pi_1\pi_2 + (1-\pi_1)(1-\pi_2) + 2\rho\sigma_1\sigma_2 \tag{3}$$

This quantity does not only involve the marginal probabilities that populations $\mathcal{R}_1$ and $\mathcal{R}_2$ classify items in category 1 (π_1 and π_2) but also the variability within each population of raters (σ_1 and σ_2) and the correlation ρ.

Under the assumption of random assignment of item i by the two populations of raters ($E[P_{i,1}P_{i,2}] = E[P_{i,1}]E[P_{i,2}]$), the *mean probability of agreement expected by chance* is simply the product of the marginal probabilities, namely

$$\Pi_E = \pi_1\pi_2 + (1-\pi_1)(1-\pi_2) \tag{4}$$

It is seen that this quantity can be obtained by setting the correlation coefficient ρ equal to 0 in Equation 3, or equivalently by setting either σ_1^2 and/or σ_2^2 equal to 0.

The agreement index between the two populations of raters is then defined in a kappa-like way, namely

$$\kappa = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E} \tag{5}$$

where $\Pi_M = max(\Pi_T)$ corresponds to the *maximum attainable value of the mean probability of agreement* (Equation 3) given the existing heterogeneity in each population of raters. Thus, $\kappa = 1$ when agreement is perfect, $\kappa = 0$ when agreement is only due to chance and $\kappa < 0$ when agreement is less than one would expect by chance.

There is a need at this stage of the development to explicit the notion of "perfect agreement" ($\kappa = 1$). By definition, the two populations of raters are said to be in perfect agreement if and only if $P_{i,1} = P_{i,2} = P_i$, for all items i in $\mathcal{I}$. In other words, the two populations of raters "perfectly" agree if and only if the probability of classifying an item in a given category is the same for the two populations. Intuitively, it is obvious that if the probability of classifying item i in category 1 is different in the two populations of raters, the latter can not agree perfectly. Note that the present definition extends that of perfect agreement between two raters, namely that $X_{i,1} = X_{i,2} = X_i$ for each item i. Under the definition of perfect agreement, if we write $E(P_i) = \pi$ and $var(P_i) = \sigma^2$, we have $ICC_g = ICC = \sigma^2/\pi(1-\pi)$, $(g = 1, 2)$ and Π_M is then given by the expression

$$\Pi_M = E(\Pi_i) = 2\sigma^2 + 2\pi^2 - 2\pi + 1 = 1 - 2\pi(1-\pi)(1 - ICC) \tag{6}$$

It is seen that $\Pi_M = 1$ if the intraclass kappa coefficient is equal to 1 in both populations of raters ($ICC = 1$, i.e. perfect agreement within each population), and/or trivially if $\pi = 0$ or $\pi = 1$ (no variability in the allocation process). Note that Schouten's agreement index is given by Equation 5 where $\Pi_M = 1$.

3.2. Nominal scale ($K > 2$)

When $K > 2$, the coefficient of agreement between two independent populations of raters is defined by

$$\kappa = \frac{\sum_{j=1}^{K}(\Pi_{[j]T} - \Pi_{[j]E})}{\sum_{j=1}^{K}(\Pi_{[j]M} - \Pi_{[j]E})} = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E} \tag{7}$$

where the quantities $\Pi_{[j]T}$, $\Pi_{[j]E}$ and $\Pi_{[j]M}$ correspond to the quantities described in the dichotomous case when the nominal scale is dichotomized by grouping all categories other than category j together and Π_T, Π_E and Π_M are defined by

$$\Pi_T = \sum_{j=1}^{K} E(P_{ij,1}P_{ij,2}); \quad \Pi_E = \sum_{j=1}^{K} \pi_{j,1}\pi_{j,2}; \quad \Pi_M = \sum_{j=1}^{K} E(P_{ij}^2).$$

and extend naturally the quantities defined in the dichotomous case. Indeed, $P_{ij,g}$ denotes the probability for item i to be classified in category j $(j = 1, \cdots, K)$ by the population of raters $\mathcal{R}_g$ $(g = 1, 2)$ and is a random variable over the population of items $\mathcal{I}$. We have $P_{ij,g} = P(X_{ijr,g} = 1|i)$ where the binary random variable $X_{ijr,g}$ is equal to 1 if rater r of population $\mathcal{R}_g$ classifies item i in category j and $\sum_{j=1}^{K} P_{ij,g} = 1$. Over the population of items $\mathcal{I}$, $E(P_{ij,g}) = \pi_{j,g}$ $(g = 1, 2)$. The equivalence of the two expressions in Equation 7 is proven in Appendix 1. The two populations of raters are defined to be in perfect agreement if and only if $P_{ij,1} = P_{ij,2} = P_{ij}$ for all items i in $\mathcal{I}$ $(j = 1, \cdots, K)$, extending the definition of the dichotomous case.

3.3. Ordinal scale (K > 2)

A weighted version of the agreement index between two populations of raters, accounting for the fact that some disagreements may be more important than others, is defined in the same way as the weighted kappa coefficient (Cohen, 1968). We have

$$\kappa_W = \frac{\Pi_{T,W} - \Pi_{E,W}}{\Pi_{M,W} - \Pi_{E,W}} \tag{8}$$

where

$$\Pi_{T,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E(P_{ij,1} P_{ik,2}), \tag{9}$$

$$\Pi_{E,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} \pi_{j,1} \pi_{k,2}, \tag{10}$$

$$\Pi_{M,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E(P_{ij} P_{ik}). \tag{11}$$

Classically, the weights w_{jk} are defined such that $0 \leq w_{jk} \leq 1$, $(j \neq k \in 1, \cdots, K)$ and $w_{jj} = 1$ $(j = 1, \cdots, K)$. The unweighted agreement index κ (see Equation 7) is obtained by using the weighting scheme $w_{jk} = 1$ if $j = k$ and $w_{jk} = 0$ otherwise $(j \neq k \in 1, \cdots, K)$.

4. Estimation of the parameters

Consider a random sample of N items from $\mathcal{I}$, a random sample of R_1 raters from $\mathcal{R}_1$ (group G_1) and a random sample of R_2 raters from $\mathcal{R}_2$ (group G_2).

4.1. Dichotomous scale (K = 2)

Suppose that $x_{ir,g}$ denote the observed values of the random variables $X_{ir,g}$ defined in section 3.1 $(i = 1, \cdots, N; r = 1, \cdots, R_g, ; g = 1, 2)$. Let

$$n_{i,g} = \sum_{r=1}^{R_g} x_{ir,g}$$

denote the number of raters of group G_g classifying item i in category 1 $(g = 1, 2)$. Then, let

$$p_{i,g} = \frac{n_{i,g}}{R_g}$$

be the corresponding proportions $(i = 1, \cdots, N; j = 1, \cdots, K; g = 1, 2)$.

At the population level, the *mean agreement over the population of items* $\mathcal{I}$ between the two populations of raters, Π_T, is estimated by the *observed proportion of agreement*

$$\widehat{\Pi}_T = p_o = \frac{1}{N}\sum_{i=1}^{N}[p_{i,1}p_{i,2} + (1 - p_{i,1})(1 - p_{i,2})]. \tag{12}$$

Likewise, the *mean probability of agreement expected by chance*, Π_E, is estimated by the *proportion of agreement expected by chance*

$$\widehat{\Pi}_E = p_e = p_1 p_2 + (1 - p_1)(1 - p_2) \tag{13}$$

where $p_g = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} p_{i,g}$ $(g = 1, 2)$.

The agreement index between the two populations of raters is then estimated by

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e} \tag{14}$$

where p_m corresponds to the maximum possible proportion of agreement derived from the samples. Indeed, recall that Π_M is obtained when $P_{i,1} = P_{i,2} = P_i$ and corresponds to the maximum expected agreement over the population of items. Thus, given the observed data, the maximum observed proportion of agreement can be obtained when $p_i = p_{i,g}$ $(g = 1, 2)$, leading to $p_o = p_{i,g}^2 + (1 - p_{i,g})^2$. Since $p_{i,1}p_{i,2} + (1 - p_{i,1})(1 - p_{i,2}) \leq max_g[p_{i,g}^2 + (1 - p_{i,g})^2]$ for each item i, it follows that

$$\widehat{\Pi}_M = p_m = \frac{1}{N}\sum_{i=1}^{N} max_g[p_{i,g}^2 + (1 - p_{i,g})^2]. \tag{15}$$

It is seen that if $p_{i,1} = p_{i,2}$ $(i = 1, \cdots, N)$, $p_o = p_m$ and $\hat{\kappa} = 1$.

4.2. Nominal case ($K > 2$)

Let $x_{ijr,g}$ denote the observed values of the random variables $X_{ijr,g}$ equal to 1 if rater r $(r = 1, \cdots, R_g)$ of population $\mathcal{R}_g$ $(g = 1, 2)$ classifies item i $(i = 1, \cdots, N)$ in category j

Table 2.

Two-way classification table of the N items by the two groups of raters on a K-categorical scale

		G_2					
	Category	1	$\dots$	j	$\dots$	K	Total
	1	c_{11}	$\dots$	c_{1j}	$\dots$	c_{1K}	$c_{1.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G_1	j	c_{j1}	$\dots$	c_{jj}	$\dots$	c_{jK}	$c_{j.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	K	c_{K1}	$\dots$	c_{Kj}	$\dots$	c_{KK}	$c_{K.}$
	Total	$c_{.1}$	$\dots$	$c_{.j}$	$\dots$	$c_{.K}$	1

$(j = 1, \cdots, K)$. The assessment of the N items by the two groups of raters can be conveniently summarized in a two-way classification table as seen in Table 2. Let

$$n_{ij,g} = \sum_{r=1}^{R_g} x_{ijr,g}$$

denote the number of raters of group G_g classifying item i in category j $(g = 1, 2)$. Then, let

$$p_{ij,g} = \frac{n_{ij,g}}{R_g}$$

be the corresponding proportions $(i = 1, \cdots, N; j = 1, \cdots, K; g = 1, 2)$. We have $\sum_{j=1}^{K} p_{ij,g} = 1$, $(i = 1, \cdots, N; g = 1, 2)$. Finally, let

$$c_{jk} = \frac{1}{N} \sum_{i=1}^{N} p_{ij,1} p_{ik,2} \; (j, k = 1, \cdots, K).$$

The quantities c_{jk} estimate the joint probability that populations $\mathcal{R}_1$ and $\mathcal{R}_2$ classify a randomly selected item i in category j and k, respectively $(c_{jk} = E(\widehat{P_{ij,1} P_{ik,2}}); j, k = 1, \cdots, K)$. A $K \times K$ matrix can then be derived from the original data (see Table 2).

The *mean probability of agreement* between the two populations of raters, Π_T, is estimated by

$$\widehat{\Pi}_T = p_o = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij,1} p_{ij,2} = \sum_{j=1}^{K} c_{jj} \qquad (16)$$

and the *mean probability of agreement expected by chance*, Π_E, is estimated by

$$\widehat{\Pi}_E = p_e = \sum_{j=1}^{K} p_{j,1}p_{j,2} = \sum_{j=1}^{K} c_{j.}c_{.j} \tag{17}$$

where $p_{j,g} = \dfrac{1}{N}\sum_{i=1}^{N} p_{ij,g}$.

The agreement index between the two populations of raters is then estimated as before by

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e} \tag{18}$$

where

$$p_m = \frac{1}{N}\sum_{i=1}^{N} max(\sum_{j=1}^{K} p_{ij,1}^2, \sum_{j=1}^{K} p_{ij,2}^2) \tag{19}$$

is the maximum possible proportion of agreement derived from the data, obtained by extending the argument developed for the dichotomous case. Note that when there is only one rater in each group of raters ($R_1 = R_2 = 1$), the agreement coefficient $\hat{\kappa}$ merely reduces to Cohen's κ coefficient (Cohen, 1960)

4.3. Ordinal scale ($K > 2$)

The weighted version of the agreement index is estimated in exactly the same way, namely

$$\widehat{\kappa}_W = \frac{p_{o,W} - p_{e,W}}{p_{m,W} - p_{e,W}} \tag{20}$$

with

$$\widehat{\Pi}_{T,W} = p_{o,w} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}p_{ij,1}p_{ik,2} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}c_{jk}, \tag{21}$$

$$\widehat{\Pi}_{E,W} = p_{e,w} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}p_{j,1}p_{k,2} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}c_{j.}c_{.k} \tag{22}$$

and

$$\widehat{\Pi}_{M,W} = p_{m,W} = \frac{1}{N}\sum_{i=1}^{N} max(\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}p_{ij,1}p_{ik,1}, \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}p_{ij,2}p_{ik,2}) \tag{23}$$

5. Consensus approach

In the theoretical framework of this paper, we have attempted to describe the consensus approach in a more formal way.

Table 3.

Expected probabilities of the classification of the two populations of raters over the sub-population $\mathcal{I}_C$ of items where a consensus exists

		$\mathcal{R}_2$		
		0	1	
$\mathcal{R}_1$	0	$E[(1-Z_{i,1})(1-Z_{i,2})]$ $(1-\phi_1)(1-\phi_2)+\rho'\sigma'_1\sigma'_2$	$E[(1-Z_{i,1})Z_{i,2}]$ $(1-\phi_1)\phi_2-\rho'\sigma'_1\sigma'_2$	$1-\phi_1$
	1	$E[Z_{i,1}(1-Z_{i,2})]$ $\phi_1(1-\phi_2)-\rho'\sigma'_1\sigma'_2$	$E(Z_{i,1}Z_{i,2})$ $\phi_1\phi_2+\rho'\sigma'_1\sigma'_2$	ϕ_1
		$1-\phi_2$	ϕ_2	1

5.1. Dichotomous scale (K=2)

Let $\mathcal{I}_C$ denote the sub-population of items on which a consensus (C) is always possible. In $\mathcal{I}_C$, consider the random variable $Z_{i,g}$ such that $Z_{i,g}=1$ if there is a consensus on category 1 for item i in the population $\mathcal{R}_g$ and $Z_{i,g}=0$ otherwise. The agreement index based on the consensus method then reduces to the case of two raters, the consensus defining a single rater in each group. Then, over $\mathcal{I}_C$, let $E(Z_{i,g})=\phi_g$ and $var(Z_{i,g})=\sigma_g'^2=\phi_g(1-\phi_g)$. If ρ' denotes the correlation coefficient between $Z_{i,1}$ and $Z_{i,2}$, we have the following representation of the expected probabilities between the two consensus values (Table 3). The agreement between the two populations of raters on item i based on the consensus, denoted Π_{iC}, is defined by

$$\Pi_{iC}=Z_{i,1}Z_{i,2}+(1-Z_{i,1})(1-Z_{i,2}) \tag{24}$$

Thus,

$$E(\Pi_{iC})=\Pi_{TC}=\phi_1\phi_2+(1-\phi_1)(1-\phi_2)+2\rho'\sigma'_1\sigma'_2 \tag{25}$$

The agreement expected by chance is defined by

$$\Pi_{EC}=\phi_1\phi_2+(1-\phi_1)(1-\phi_2) \tag{26}$$

and perfect agreement is achieved when $Z_{i,1}=Z_{i,2}$, for all items in $\mathcal{I}_C$, leading to

$$E(\Pi_{iC})=\Pi_{MC}=1$$

Therefore, the agreement coefficient between the two populations of raters is defined by

$$\kappa_C = \frac{\Pi_{TC} - \Pi_{EC}}{1 - \Pi_{EC}} \tag{27}$$

5.2. Nominal scale ($K > 2$)

Consider the random variable $Z_{ij,g}$ such that $Z_{ij,g} = 1$ if there is a consensus on category j for item i in population $\mathcal{R}_g$ and $Z_{ij,g} = 0$ otherwise. Then, over $\mathcal{I}_C$, let $E(Z_{ij,g}) = \phi_{j,g}$. In the same way as before,

$$\kappa_C = \frac{\sum_{j=1}^{K}(\Pi_{[j]TC} - \Pi_{[j]EC})}{\sum_{j=1}^{K}(\Pi_{[j]MC} - \Pi_{[j]EC})} = \frac{\Pi_{TC} - \Pi_{EC}}{\Pi_{MC} - \Pi_{EC}} \tag{28}$$

where $\Pi_{[j]TC}$, $\Pi_{[j]EC}$ and $\Pi_{[j]MC}$ correspond to the quantities described in the dichotomous case when the nominal scale is dichotomized by grouping all categories other than category j together. The quantities Π_{TC}, Π_{EC} and Π_{MC} are defined respectively by

$$\Pi_{TC} = \sum_{j=1}^{K} E(Z_{ij,1}Z_{ij,2});\ \Pi_{EC} = \sum_{j=1}^{K} \phi_{j,1}\phi_{j,2};\ \Pi_{MC} = 1.$$

5.3. Ordinal scale ($K > 2$)

The weighted version of the consensus approach can also be derived in the same way as before by introducing weights in the expression of Π_{TC}, Π_{EC} and Π_{MC}.

$$\Pi_{T,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}E(Z_{ij,1}Z_{ik,2}); \tag{29}$$

$$\Pi_{E,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}\phi_{j,1}\phi_{k,2}; \tag{30}$$

$$\Pi_{M,WC} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk}E(Z_{ij}Z_{ik}) = 1 \tag{31}$$

leading to

$$\kappa_{C,W} = \frac{\Pi_{T,WC} - \Pi_{E,WC}}{1 - \Pi_{E,WC}} \tag{32}$$

5.4. Remark

The consensus approach is equivalent to the new agreement index if and only if $R_1 = R_2 = 1$ or if and only if a consensus is always possible for each item in both populations of raters ($\mathcal{I}_C = \mathcal{I}$) and there is perfect agreement in both populations of raters ($P_{ij,1} = P_{ij,2} = P_{ij}$, $\forall i$). It can also be shown that with the additional assumption $ICC_1 = ICC_2 = 1$ (perfect agreement in each population of raters), the agreement index κ is algebraically equivalent to the inter-cluster agreement index introduced by Schouten (1982).

5.5. Estimation of the parameters

Consider again a random sample of R_1 raters from $\mathcal{R}_1$, a random sample of R_2 raters from $\mathcal{R}_2$ and a random sample of N items from $\mathcal{I}$. Let N_C $(\leq N)$ denote the number of items where a consensus exist in each group. Suppose that $z_{ij,g}$ denotes the observed values of the random variables $Z_{ij,g}$ $(i = 1, \cdots, N_C; j = 1, \cdots, K; g = 1, 2)$ defined in the previous section. The assessment of the N_C items on which the two groups of raters can determine a consensus can be conveniently

$$d_{jk} = \frac{1}{N_C} \sum_{i=1}^{N_C} z_{ij,1} z_{ik,2} \ (j, k = 1, \cdots, K).$$

Similarly to what was done in Section 4, the *observed weighted agreement* between the two groups of raters is obtained by

$$\widehat{\Pi}_{T,WC} = p_{o,WC} = \frac{1}{N_C} \sum_{i=1}^{N_C} \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} z_{ij,1} z_{ik,2} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} d_{jk} \tag{33}$$

and the *agreement expected by chance* by the expression

$$\widehat{\Pi}_{E,WC} = p_{e,WC} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} z_{j,1} z_{k,2} = \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} d_{j.} d_{.k} \tag{34}$$

where $z_{j,g} = \frac{1}{N_C} \sum_{i=1}^{N_C} z_{ij,g}$, $(g = 1, 2)$ leading to the weighted agreement coefficient

$$\hat{\kappa}_{C,W} = \frac{p_{o,WC} - p_{e,WC}}{1 - p_{e,WC}}. \tag{35}$$

6. Sampling variance

The Jackknife method (Efron & Tibshirani, 1993) was used to determine the sampling variance of the agreement indexes. Suppose that the agreement between two independent groups of raters was estimated from a random sample of N observations. Let $\widehat{\kappa}_N$ denotes the agreement index between the two groups of raters. Let $\widehat{\kappa}_{N-1}^{(i)}$ denotes the estimated agreement coefficient when item i is deleted. These quantities are used to determine the pseudo-values

$$\widehat{\kappa}_{N,i} = N\widehat{\kappa}_N - (N-1)\widehat{\kappa}_{N-1}^{(i)}$$

The Jackknife estimator of the agreement index is then defined by

$$\widetilde{\kappa_N} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\kappa}_{N,i}$$

with variance

$$var(\widetilde{\kappa}_N) = \frac{1}{N}\left\{\frac{1}{N-1}\sum_{i=1}^{N}(\widehat{\kappa}_{N,i} - \widehat{\kappa}_N)^2\right\}$$

The bias of the Jackknife estimator is estimated by

$$Bias(\widetilde{\kappa}_N) = (N-1)\left\{\widetilde{\kappa}_N - \widehat{\kappa}_N\right\}.$$

7. Example: Script Concordance Test

The script concordance test (SCT) (Charlin & al., 2002) is used in medical education to score physicians or medical students in their ability to solve clinical situations as compared to answers given by experts. The test consists of a number of items to be evaluated on a 5-point Likert scale. Each item represents a clinical situation (called an "assumption") likely to be encountered in the physician's practice. The situation has to be unclear, even for an expert. The task of the subjects being evaluated is to consider the effect of new information on the assumption to solve the situation. In this respect, they have to choose between the following proposals: (-2) The assumption is practically eliminated; (-1) The assumption becomes less likely; (0) The information has no effect on the assumption; (+1) The assumption becomes more likely; (+2) The assumption is virtually the only possible one. The present research project has been motivated by the problem of finding the overall degree of agreement between the responses given to the SCT by the candidates and those given by a panel of medical experts.

During the period 2003-2005, the SCT was proposed to medical students training in "general practice" (Vanbelle & al., 2007). The SCT consisted of 34 items relating possible situations (assumptions) encountered in general practice. A total of 39 students passed the

Table 4.

Two-way classification table of the 34 items of the Script Concordance Test (SCT) by the group of 11 medical experts and by the group of 39 medical students using a 5-point Likert scale

		Medical experts					
		(-2)	(-1)	(0)	(1)	(2)	Total
Medical students	(-2)	0.077	0.054	0.028	0.009	0.002	0.170
	(-1)	0.036	0.067	0.066	0.033	0.012	0.214
	(0)	0.022	0.053	0.187	0.062	0.013	0.337
	(1)	0.013	0.026	0.069	0.090	0.025	0.223
	(2)	0.005	0.009	0.013	0.020	0.010	0.057
	Total	0.153	0.209	0.363	0.214	0.057	1

(-2) The assumption is practically eliminated; (-1) The assumption becomes less likely;

(0) The information has no effect on the assumption; (+1) The assumption becomes more likely

(+2) The assumption is practically the only possible

test. Their responses were confronted to those of a panel of 11 experts. Thus, in the present example, $R_1 = 11$, $R_2 = 39$, $N = 34$ and $K = 5$. The cross-classification matrix $(c_{jk}, j, k = 1, \cdots, 5)$ between the group of medical students and the group of experts is given in Table 4. Since the scale is ordinal, weighted agreement indexes were calculated using the quadratic weighting scheme $(w_{jk} = 1 - (|k-j|/4)^2$, $k, j = -2, \cdots, 2)$ (Fleiss & Cohen, 1973). On the basis of the study material, we found that the observed proportion of agreement, the proportion of agreement expected by chance and the maximum proportion of agreement were respectively $p_o = 0.80$, $p_e = 0.69$ and $p_m = 0.84$, yielding a weighted agreement index $\hat{\kappa}_W = (0.80 - 0.69)/(0.84 - 0.69) = 0.72$. In Table 5, $\hat{\kappa}_{C,W1}$ corresponds to the consensus method using the majority rule and $\hat{\kappa}_{C,W2}$ to the 50% rule (Equation 35), while $\hat{\kappa}_{S,W}$ is the agreement coefficient derived by Schouten (1982). It should be noted that there were 2 items without consensus for the majority rule and 16 for the 50% rule. When calculating the mean (± SD) of weighted kappa coefficients for all possible pairs of raters (429 pairs) between the two groups, we obtained 0.35 ± 0.06, a value similar to Schouten's index. The intraclass correlation coefficient was 0.22 ± 0.04 in the group of experts and 0.29 ± 0.03 in the group of students, reflecting a substantial heterogeneity in both groups.

Table 5.

Weighted agreement indexes between the group of 11 experts and the group of 39 students for the Script Concordance Test (SCT) with 34 items obtained by four different methods with quadratic weighting scheme.

Method	Coefficient	N	p_o	p_e	p_m	$\hat{\kappa}$	SE(κ)
Proposed	$\hat{\kappa}_W$	34	0.80	0.69	0.84	0.72	0.049
Consensus (majority)	$\hat{\kappa}_{C,W1}$	32	0.88	0.71	1	0.60	0.11
Consensus (50%)	$\hat{\kappa}_{C,W2}$	18	0.93	0.60	1	0.82	0.11
Schouten	$\hat{\kappa}_{S,W}$	34	0.80	0.69	1	0.35	0.049

8. Discussion

Cohen's kappa coefficient (Cohen, 1960) is widely used to measure agreement between two raters judging items on a categorical scale. Weighted (Cohen, 1968) and intraclass (Kraemer, 1979) versions of the coefficient were also proposed. Further, the method was extended to several raters (Fleiss, 1981). The modelling of the kappa coefficient with respect to categorical and continuous covariates has been extensively investigated in recent years (Williamson & al. (2000); Lipsitz & al. (2001); Barnath & Williamson (2002)), hence providing a comprehensive analytical approach to this important concept.

The problem of assessing the agreement between two groups of raters is not new. Applications are numerous (e.g., van Hoeij & al. (2004); Raine & al. (2004)) and a variety of methods has been proposed over the years to deal with this problem. Several recent articles from the applied field (e.g. Kraemer & al., 2004), however, while emphasing the importance and relevance of the problem, claim that existing solutions are not quite appropriate and that there is a need for novel and improved methods.

The usual way to solve the problem of agreement between two groups of raters is to define a consensus in each group and to quantify the agreement between them. The problem is then reduced to the case of computing Cohen's kappa agreement coefficient between two raters on a categorical scale. The rule of consensus may be defined as choosing for each item the modal (or majority) category or the category whose frequency exceeds a given percentage (e.g. 50% or 80%) in each group of raters. The consensus method, however, has serious limitations that weaken its use in practice. Indeed, a consensus is not always possible for all items (as illustrated by the SCT data) resulting in a loss of items and hence of statistical precision. The variability of the responses within each group of raters is completely ignored and the strength of the consensus is not really reflected. Further, the conclusions can be highly dependent on which definition is used for the consensus (Kraemer & al., 2004). Moreover, since items without consensus (i.e., with high variability among the raters) are generally discarded from the analysis, the results obtained are prone to bias and over-optimistic estimation (see SCT example). Another natural method for assessing the concordance between two sets of raters consists in calculating the mean kappa coefficient between all possible pairs of raters composed by one rater of each group. As seen in the SCT example, this approach gives a value similar to the index developed by Schouten (1982) in the context of hierarchical clustering of raters within a single population of raters.

The agreement between two groups of raters raises the basic question of what it meant by "perfect agreement" between two groups. While this issue is meaningless in the case of two raters (they agree or they don't agree), it becomes critical at the group level agreement. The consensus method is one way to circumvent the difficulty and the mean of all pairwise kappa coefficients in another way. Schouten (1982) eluded the problem by defining perfect agreement between two groups as the situation where all raters of each group perfectly agree on all items, quite an extreme assumption. The novelty of the method derived in this paper is that it rests on a less stringent definition of perfect agreement in a population-based context. Specifically, two populations of raters are defined to be in perfect agreement (kappa coefficient equal to 1) if they have the same probability of classifying each item on the K-categorical scale. With this definition in mind, it does not really matter which raters agree or don't agree for a given item within each population, as long as the proportions in the two populations are equal. Each population is viewed as a global entity with its own heterogeneity and there is no direct interest in the agreement of individual raters within or between populations. Actually, it is quite possible that the two populations perfectly agree while a substantial part of raters disagree with each other in their own population and with some raters in the other population. As a consequence of the definition of perfect agreement, the maximum attainable proportion of agreement between the two populations (at least in the dichotomous case) can be expressed as an analytical function of two factors, the intraclass correlation coefficient within each population and the overall marginal probabilities of classifying the items. By setting the intraclass correlation coefficient equal to 1, it turns out that our approach rejoins Schouten's assumption of perfect agreement, which can therefore be regarded as a special (extreme) case of our general definition. As illustrated on the SCT data, the difference between Schouten's approach and ours can be marked ($\hat{\kappa} = 0.72$ and 0.35, respectively). This is due to the fact that both groups of raters show a high variability in their responses (the ICC was 0.22 ± 0.04 in the group of experts and 0.29 ± 0.03 in the group of students, respectively). The present method allows for prefect agreement in presence of group heterogeneity while Schouten's approach does not. Schouten's index, however, can be derived directly from the $K \times K$ contingency table of joint probabilities estimates (see Table 3), whereas this is not possible with the proposed approach because the definition of perfect agreement requires the raw original data to be available to compute the maximum attainable value.

The new agreement index is also superior to the consensus approach (a method that we tried to formalize more theoretically) in the sense that it takes into account the variability among raters in each population and it incorporates always all items to be allocated. An intraclass and weighted versions were also proposed. If there is only one rater in each group, all coefficients envisaged here reduce to Cohen's kappa coefficient. Recently, Vanbelle and Albert (2008) envisaged the agreement between a single rater and a group of raters, a situation which may be regarded as a special case of the present one but which raises specific problems in practice.

The estimation of the kappa coefficient is fairly straightforward, although the calculation of the maximum proportion of agreement requires particular attention. As for the sampling variability aspects, we suggested to use the Jackknife method rather than by asymptotic formulas.

In conclusion, the index proposed in this paper measures the overall agreement between two independent groups of raters, taking into account the within group heterogeneity. The method is a natural extension of Cohen's kappa coefficient and demonstrates similar properties.

Table 6.

2×2 table cross-classifying the two populations of raters with respect to a nominal scale, obtained when grouping all categories other than category $[j]$ together

		$\mathcal{R}_2$		
		[j]	Other	
$\mathcal{R}_1$	[j]	$E[P_{ij,1}P_{ij,2}]$	$E[P_{ij,1}(1-P_{ij,2})]$	$\pi_{j,1}$
	Other	$E[(1-P_{ij,1})P_{ij,2}]$	$E[(1-P_{ij,1})(1-P_{ij,2})]$	$1-\pi_{j,1}$
		$\pi_{j,2}$	$1-\pi_{j,2}$	1

9. Appendix

Equivalence 1. We have

$$\kappa = \frac{\sum_{j=1}^{K}(\Pi_{[j]T}-\Pi_{[j]E})}{\sum_{j=1}^{K}(\Pi_{[j]M}-\Pi_{[j]E})} = \frac{\Pi_T-\Pi_E}{\Pi_M-\Pi_E}$$

where the quantities $\Pi_{[j]T}$, $\Pi_{[j]E}$ and $\Pi_{[j]M}$ correspond to the quantities described in the dichotomous case when the nominal scale is dichotomized by grouping all categories other than category j together and Π_T, Π_E and Π_M are defined by

$$\Pi_T = \sum_{j=1}^{K} E(P_{ij,1}P_{ij,2}); \quad \Pi_E = \sum_{j=1}^{K} \pi_{j,1}\pi_{j,2}; \quad \Pi_M = \sum_{j=1}^{K} E(P_{ij}^2).$$

Proof. Indeed, when grouping all categories other than $[j]$ together, a 2×2 table cross-classifying populations of raters $\mathcal{R}_1$ and $\mathcal{R}_2$ with respect to category j of the nominal scale can be constructed ($j=1,\cdots,K$) (Table 6).

Thus,

$$\begin{aligned}
\sum_{j=1}^{K}\Pi_{[j]T} &= \sum_{j=1}^{K} E[P_{ij,1}P_{ij,2}+(1-P_{ij,1})(1-P_{ij,2})]) \\
&= E(2\sum_{j=1}^{K} P_{ij,1}P_{ij,2}+\sum_{j=1}^{K}1-\sum_{j=1}^{K}P_{ij,1}-\sum_{j=1}^{K}P_{ij,2}) \\
&= 2E(\sum_{j=1}^{K} P_{ij,1}P_{ij,2})+K-2 \\
&= 2\Pi_T+K-2
\end{aligned}$$

Likewise, it is easily seen that

$$\sum_{j=1}^{K} \Pi_{[j]E} = \Pi_E + K - 2 \text{ and } \sum_{j=1}^{K} \Pi_{[j]M} = \Pi_M + K - 2.$$

It follows immediately that

$$\kappa = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E}.$$

References

Barnhart, H.X., & Williamson, J.M. (2002). Weighted least squares approach for comparing correlated kappa, *Biometrics*, *58*, 1012–1019

Bland, A.C., Kreiter, C.D., & Gordon, J.A., (2005). The psychometric properties of five scoring methods applied to the Script Concordance Test. *Academic Medicine*, *80*, 395–399.

Charlin, B., Gagnon, R., Sibert, L., & Van der Vleuten, C. (2002). Le test de concordance de script: un instrument d'évaluation du raisonnement clinique. *Pédagogie Médicale*, *3*, 135–144.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological bulletin*, *70*, 213–220.

Efron, B. & Tibshirani, R.J. (1993). An introduction to the bootstrap. *Chapman and Hall, New York.*

Feigin, P.D., & Alvo, M. (1986). Intergroup Diversity and Concordance for Ranking Data: An Approach via Metrics for Permutations. *The Annals of Statistics*, *14*, 691–707

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*, John Wiley, New York, 2nd edition.

Fleiss, J.L. and J. Cohen (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability, *Educational and psychological measurement*, *33*, 613–619.

Hollander, M., & Sethuraman, J. (1978). Testing for agreement between two groups of judges. *Biometrika*, *65*, 403–411.

Kraemer, H.C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, *44*, 461–472.

Kraemer, H.C. (1981). Intergroup concordance: Definition and estimation. *Biometrika*, *68* 641–646.

Kraemer, H. C., Vyjeyanthi, S.P., & Noda, A. (2004). Agreement Statistics. In R.B. D'Agostino, *Tutorial in Biostatistics vol 1.* (pp. 85–105.), John Wiley and Sons.

Lipsitz, S.R., Williamson, J., Klar, N., Ibrahim, J. & Parzen, M. (2001) A simple method for estimating a regression model for κ between a pair of raters, *Journal of the Royal Statistical Society series A*, *164*, 449–465

Raine, R., Sanderson, C., Hutchings, A., Carter, S., Larking, K., & Black, N. (2004). An experimental study of determinants of group judgments in clinical guideline development. *Lancet*, *364*, 429–437.

Schouten H.J.A. (1982). Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica*, *36*, 45–61.

Schucany, W.R., & Frawley, W.H. (1973). A rank test for two group concordance. *Psychometrika*, *38*, 249–258.

van Hoeij, M.J., Haarhuis, J.C., Wierstra ,R.F., & van Beukelen, P. (2004). Developing a classification tool based on Bloom's taxonomy to assess the cognitive level of short essay questions. *Journal of Veterinary Medical Education*, *31*, 261–267.

Vanbelle, S., Massart, V., Giet, G., & Albert, A. (2007). Test de concordance de script: un nouveau mode d'établissement des scores limitant l'effet du hasard. *Pédagogie Médicale*, *8*, 71–81.

Vanbelle, S., & Albert, A. (2008). Agreement between an isolated rater and a group of raters *Statistica Neerlandica*, in press.

Williamson, J.M., Lipsitz, S.R., & Manatunga, A.K. (2000). Modeling kappa for measuring dependent categorical agreement data, *Biostatistics*, *1*, 191–202