Statistica Neerlandica (2009) Vol. 63, nr. 1, pp. 82–100
doi:10.1111/j.1467-9574.2008.00412.x

Agreement between an isolated rater and a group of raters

S. Vanbelle* and A. Albert

Medical Informatics and Biostatistics, University of Liège, CHU Sart Tilman (B23), 4000 Liège, Belgium

The agreement between two raters judging items on a categorical scale is traditionally assessed by Cohen's kappa coefficient. We introduce a new coefficient for quantifying the degree of agreement between an isolated rater and a group of raters on a nominal or ordinal scale. The group of raters is regarded as a whole, a reference or gold-standard group with its own heterogeneity. The coefficient, defined on a population-based model, requires a specific definition of the concept of perfect agreement. It has the same properties as Cohen's kappa coefficient and reduces to the latter when there is only one rater in the group. The new approach overcomes the problem of consensus within the group of raters and generalizes Schouten's index. The method is illustrated on published syphilis data and on data collected from a study assessing the ability of medical students in diagnostic reasoning when compared with expert knowledge.

Keywords and Phrases: kappa coefficient, nominal scale, ordinal scale, expert group.

1 Introduction

Cohen (1960) introduced the kappa coefficient $\hat{\kappa} = (p_o - p_e)/(1 - p_e)$ to quantify the agreement between two raters classifying items on a categorical scale. He corrected the proportion of items with concordant classification (p_o) for the proportion of concordant pairs expected by chance (p_e) and standardized the quantity to obtain 1 in case of a perfect agreement between the two raters and 0 when the raters agree by chance. There are situations where agreement is searched between an isolated rater and a group of raters, regarded as a whole, a reference, expert or gold-standard group, in which all raters may not perfectly agree with each other. For example, each of a series of candidates may be assessed against a group of experts with the purpose of evaluating their knowledge and of classifying the candidates. This is a frequent exercise in education or in competence examinations. In the context of accreditation, a routine laboratory may have to reach a pre-defined level of agreement when challenged against a set of reference laboratories for a number of test

*sophie.vanbelle@ulg.ac.be

© 2009 The Authors. Journal compilation © 2009 VVS.
Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

specimens. This process has to account for the fact that the reference laboratories exhibit themselves analytical variability and do not necessarily agree with each other. The traditional approach to solve the problem is to determine a consensus in the group of raters and to measure the agreement between the isolated rater and the consensus in the group (Landis and Koch, 1977; Soeken and Prescott, 1986; Salerno, Alguire and Waxman, 2003). Thus, the so-called 'consensus method' reduces the problem of computing the classical Cohen's kappa coefficient. Consensus may be defined as the category chosen by a given proportion of raters in the group (e.g. Ruperto *et al.*, 2006, defined consensus as the category chosen by at least 80% of the raters in the group) or the category the most frequently chosen by the raters in the group (Kalant *et al.*, 2000; Smith *et al.*, 2003). In both cases, however, the problem of handling items without consensus in the group arises. Ruperto *et al.* (2006) discarded all items without consensus from the analysis, while Kalant *et al.* (2000) and Smith *et al.* (2003) did not encounter the problem. The method consisting in reducing the judgements made by a group of raters into a consensus decision was criticized by Eckstein *et al.* (1998), Salerno *et al.* (2003) and Miller *et al.* (2004). Eckstein *et al.* (1998) studied the bias that may result from removing items without consensus, while Salerno *et al.* (2003) argued that the dispersion likely to occur in the classifications made by the raters in the group may not be reflected in the consensus. Finally, Miller *et al.* (2004) showed that different conclusions may be obtained by using different rules of consensus. Williams (1976) developed a measure for comparing the joint agreement of several raters with another rater without determining a consensus in the group of raters. Specifically, he compared the mean proportion of concordant items between the isolated rater and each rater in the group with the mean proportion of concordant items between all possible pairs of raters among the group. The ratio derived, known as 'Williams' index', is compared with the value of 1. Unfortunately, Williams' index does neither account for agreement due to chance nor measure the agreement between the isolated rater and the group of raters. In a different context, Schouten (1982) described a hierarchical clustering method based on pairwise weighted agreement measures (referred hereafter as 'Schouten's agreement index') to identify homogeneous subgroups among a group of raters classifying items on a nominal or ordinal scale. Finally, Light (1971) investigated the reverse problem of comparing the joint agreement of several raters with a gold standard. He derived a statistic based on the proportion of concordant pairs obtained between each rater in the group and the gold standard (the isolated rater). As for Williams' index, Light's method does not actually quantify the agreement between the gold standard and the group of raters.

In the present study, a novel coefficient is proposed for quantifying the agreement between an isolated rater and a group of raters, considered as a well-defined entity with its own heterogeneity. This coefficient overcomes the problems of consensus by capturing the variability within the group of raters. It generalizes the approach of Schouten (1982) and possesses the same properties as Cohen's kappa coefficient. In Section 2, we briefly recall the intraclass kappa coefficient (ICC) quantifying the

© 2009 The Authors. Journal compilation © 2009 VVS.

agreement within a group of raters. In Section 3, we introduce the new agreement coefficient from a population-based perspective, not only for binary or multinomial scales but also for ordinal scales. Sections 4 and 5 are devoted to the estimation of the agreement index and its asymptotic sampling variance, respectively. Section 6 looks into consensus estimation, while in Section 7, the method is illustrated on real-life examples. The paper closes with a discussion in Section 8. Some proofs and detailed calculations are appended in Section 9.

2 Agreement between several raters

Consider a population $\mathcal{I}$ of items and a population $\mathcal{R}$ of raters. Suppose that the items have to be classified into two categories ($K=2$) by the raters. Consider a randomly selected rater r from population $\mathcal{R}$ and a randomly selected item i from population $\mathcal{I}$. Let $X_{i,r}$ be the random variable such that $X_{i,r}=1$ if rater r classifies item i in category 1 and $X_{i,r}=0$ otherwise. For each item i, $E(X_{i,r}\,|\,i)=P(X_{i,r}=1)=P_i$ over the population of raters and $\text{var}(X_{i,r}\,|\,i)=P_i(1-P_i)$. Then, over the population of items, $E(P_i)=E[E(X_{i,r}\,|\,i)]=\pi$ and $\text{var}(P_i)=\sigma^2$. The agreement in the population of raters is classically quantified by the ICC (Kraemer, 1979)

$$\text{ICC}=\frac{\sigma^2}{\pi(1-\pi)}.$$

It is easily shown that $0\leq \text{ICC}\leq 1$. The value $\text{ICC}=1$ corresponds to perfect agreement within the population of raters. By contrast, $\text{ICC}=0$ when the heterogeneity of the items is not well detected by the raters or when items are homogeneous in the population (Kraemer, Periyakoil and Noda, 2002).

3 Definition of the agreement index

3.1 Binary scale ($K=2$)

Using the notation introduced in Section 2, consider an isolated rater not belonging to $\mathcal{R}$. Suppose that all raters from population $\mathcal{R}$ and the isolated rater have to classify a randomly selected item i from $\mathcal{I}$ in two categories ($K=2$). Let Y_i denote the random variable equal to 1 if the isolated rater classifies item i in category 1 and $Y_i=0$ otherwise. Over the population of items, $E(Y_i)=\pi^*$ and $\text{var}(Y_i)=\sigma^{*2}=\pi^*(1-\pi^*)$. The correlation between P_i and Y_i over $\mathcal{I}$ implies

$$\rho=\frac{E(P_i Y_i)-\pi\pi^*}{\sigma\sigma^*}.$$

Now, consider the joint probability distribution of the classification of item i made by the population of raters and the isolated rater. On a binary scale, this consists of four probabilities $(1-P_i)(1-Y_i), (1-P_i)Y_i, P_i(1-Y_i)$ and P_iY_i, respectively. For example, P_iY_i denotes the probability that the population of raters and the isolated rater both classify item i in category 1. The expectations, over the population of

© 2009 The Authors. Journal compilation © 2009 VVS.

Table 1. Expected joint and marginal probability distributions resulting from the binary classification of a randomly selected item i from the population $\mathcal{I}$ by the population of raters $\mathcal{R}$ and the isolated rater.

	Isolated rater		
$\mathcal{R}$	0	1	
0	$E[(1-P_i)(1-Y_i)]$ $(1-\pi)(1-\pi^*)+\rho\sigma\sigma^*$	$E[(1-P_i)Y_i]$ $(1-\pi)\pi^*-\rho\sigma\sigma^*$	$1-\pi$
1	$E[P_i(1-Y_i)]$ $\pi(1-\pi^*)-\rho\sigma\sigma^*$	$E[P_iY_i]$ $\pi\pi^*+\rho\sigma\sigma^*$	π
	$1-\pi^*$	π^*	1

items, of these joint probabilities can be represented in a 2×2 classification table, as displayed in Table 1.

The probability that the population of raters and the isolated rater agree on item i is given by

$$\Pi_i = P_iY_i + (1-P_i)(1-Y_i) \tag{1}$$

so that, over the population of items $\mathcal{I}$, the mean probability of agreement is given by the expression

$$\Pi_T = E(\Pi_i) = \pi\pi^* + (1-\pi)(1-\pi^*) + 2\rho\sigma\sigma^* \tag{2}$$

which corresponds to the sum of the diagonal elements in Table 1. Surprisingly, for a given level of agreement (ICC) within the population of raters, the maximum attainable value Π_T is not necessarily equal to 1 as shown below.

By definition, the population of raters and the isolated rater 'perfectly agree' when $\pi=\pi^*$ and $\rho=1$. In terms of the random variables P_i and Y_i, this is equivalent to writing (see proof in Section 9.1)

$$P_i = \pi^{**}(1-\sqrt{\text{ICC}}) + \sqrt{\text{ICC}}\,Y_i.$$

where, for convenience, π^{**} denotes the common value of $\pi=\pi^*$.

Replacing P_i in Equation 1 and taking the expectation over population $\mathcal{I}$, the maximum attainable value of Π_T is found to be

$$\Pi_M = 1 - 2\pi^{**}(1-\pi^{**})(1-\sqrt{\text{ICC}}) \tag{3}$$

This quantity turns out to be equal to 1 if and only if ICC$=1$, i.e. there is perfect agreement in the population of raters $\mathcal{R}$, or trivially, if $\pi^{**}=0$ or 1. It should be remarked at this stage that Schouten (1982), in his paper, implicitly assumed $\Pi_M=1$.

Following the results above, the coefficient of agreement between the population of raters and the isolated rater can be advantageously defined in a kappa-like manner, namely,

© 2009 The Authors. Journal compilation © 2009 VVS.

$$\kappa = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E} \tag{4}$$

with Π_T the theoretical agreement, Π_M the maximum attainable agreement and Π_E the agreement expected by chance. Π_E is the probability that the population of raters and the isolated rater agree under the independence assumption, $E(P_i Y_i) = E(P_i)E(Y_i)$. Π_E is defined by

$$\Pi_E = \pi\pi^* + (1-\pi)(1-\pi^*). \tag{5}$$

Note that $\Pi_T = \Pi_E$ (see Equations 2 and 5) in the absence of correlation between the ratings of the population of raters and of the isolated rater ($\rho = 0$) or when there is no variability in the classifications made by the population of raters ($\sigma^2 = 0$) or by the isolated rater ($\sigma^{*2} = 0$). The agreement coefficient (Equation 4) has been standardized in such a way that $\kappa = 1$ if the agreement between the isolated rater and the group of raters reaches the maximum attainable value Π_M (perfect agreement) and $\kappa = 0$ when agreement can only be explained by pure chance. Finally, observe that Equation 4 reduces to Schouten's index when $\Pi_M = 1$.

3.2 Multinomial scale ($K > 2$)

When $K > 2$, the coefficient of agreement between the population of raters and the isolated rater is defined by

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

where $\Pi_{[j]T}$, $\Pi_{[j]E}$ and $\Pi_{[j]M}$ correspond to the quantities described in the binary case ($K = 2$) when the nominal scale is dichotomized by grouping all categories other than category j together. Π_T, Π_E and Π_M are defined by

$$\Pi_T = \sum_{j=1}^{K} E[P_{ij} Y_{ij}]; \quad \Pi_E = \sum_{j=1}^{K} \pi_j \pi_j^*;$$

$$\Pi_M = \sum_{j=1}^{K} E[(\pi_j^{**} + (1-\pi_j^{**})\sqrt{\mathrm{ICC}_j}) Y_{ij}]$$

$$= \sum_{j=1}^{K} (\pi_j^{**} + \pi_j^{**}(1-\pi_j^{**})\sqrt{\mathrm{ICC}_j})$$

where P_{ij} denotes the probability for a randomly selected item i to be classified in category j ($j = 1, \ldots, K$) by the population of raters, with $E(P_{ij}) = \pi_j$. Y_{ij} denotes the random variable equal to 1 if the isolated rater classifies item i in category j ($Y_{ij} = 0$ otherwise). Finally, ICC_j denotes the ICC relative to category j ($j = 1, \ldots, K$) in the population of raters (see proof in Section 9.2).

© 2009 The Authors. Journal compilation © 2009 VVS.

The coefficient κ possesses the same properties as Cohen's kappa coefficient, $\kappa = 1$ when agreement is perfect ($\Pi_T = \Pi_M$), $\kappa = 0$ if observed agreement is equal to agreement expected by chance ($\Pi_T = \Pi_E$) and $\kappa < 0$ if observed agreement is lower than that expected by chance ($\Pi_T < \Pi_E$).

3.3 Ordinal scale ($K > 2$)

A weighted version of the agreement index can be defined in a way similar to the weighted kappa coefficient (Cohen, 1968),

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

with

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

$$\Pi_{E,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} \pi_j \pi_k^*;$$

$$\Pi_{M,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E[(\pi_j^{**} + (1-\pi_j^{**})\sqrt{\mathrm{ICC}_j}\, Y_{ij})\, Y_{ik}].$$

In general, $0 \le w_{jk} \le 1$ and $w_{kk} = 1 (j, k = 1, \ldots, K)$. Cicchetti and Allison (1971) have defined absolute weights

$$w_{jk} = 1 - \frac{|j-k|}{K-1},$$

whereas Fleiss and Cohen (1973) suggested quadratic weights

$$w_{jk} = 1 - \left(\frac{j-k}{K-1}\right)^2.$$

4 Estimation of the parameters

Consider a random sample of N items drawn from population $\mathcal{I}$. Let each item be classified independently on a K-categorical scale by a random sample (group) of R raters from population $\mathcal{R}$ and by the isolated rater.

4.1. Binary scale

Let $x_{i,r}$ designate the observed value of the random variable $X_{i,r}$, denoting the category assignment made for item i by rater r from population $\mathcal{R}$ ($i = 1, \ldots, N$; $r = 1, \ldots, R$). Then, let

© 2009 The Authors. Journal compilation © 2009 VVS.

$$n_i = \sum_{r=1}^{R} x_{i,r}$$

denote the number of times that item i is classified in category 1 by the group of raters and $p_i = n_i/R$ the corresponding proportion ($i = 1, \ldots, N$).

The ICC in the group of raters can be estimated by the expression (Fleiss, 1981)

$$\widehat{\mathrm{ICC}} = 1 - \frac{\sum_{i=1}^{N} n_i(R - n_i)}{RN(N-1)p(1-p)}$$

where p is the overall proportion of items classified in category 1 by the group of raters,

$$p = \frac{1}{N}\sum_{i=1}^{N} p_i.$$

If y_i denotes the observed value of the random variable Y_i, representing the category assignment of item i by the isolated rater, the probability that the population of raters and the isolated rater agree is estimated by the *observed proportion of agreement*,

$$p_o = \hat{\Pi}_T = \frac{1}{N}\sum_{i=1}^{N} [p_i y_i + (1 - p_i)(1 - y_i)]. \tag{6}$$

The probability of agreement expected by chance is estimated by the *proportion of agreement expected by chance*,

$$p_e = \hat{\Pi}_E = py + (1-p)(1-y)$$

where y is the proportion of items classified in category 1 by the isolated rater,

$$y = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

The degree of agreement κ between the group of raters and the isolated rater is then estimated by

$$\hat{\kappa} = \frac{p_o - p_e}{p_m - p_e}$$

where p_m corresponds to the maximum possible proportion of agreement derived from the sample. As each response y_i given by the isolated rater can only be 0 or 1, it is easily seen that for each item i, $p_i y_i + (1 - p_i)(1 - y_i) \leq \max(p_i, 1 - p_i)(i = 1, \ldots, N)$. It follows from Equation 6 that the maximum attainable proportion of agreement is given by the expression

$$p_m = \hat{\Pi}_M = \frac{1}{N}\sum_{i=1}^{N} \max(p_i, 1 - p_i).$$

© 2009 The Authors. Journal compilation © 2009 VVS.

This quantity can only be equal to 1 if $p_i = 0$ or 1 for all items ($i = 1, \ldots, N$) as assumed by Schouten.

4.2 Multinomial scale ($K > 2$)

The estimation of the parameters easily extends to the case $K > 2$. Let $x_{ij,r}$ denote the observed value of the random variable $X_{ij,r}$ equal to 1 if rater r ($r = 1, \ldots, R$) of the group classifies item i ($i = 1, \ldots, N$) in category j ($j = 1, \ldots, K$) and equal to 0 otherwise. Then, let

$$n_{ij} = \sum_{r=1}^{R} x_{ij,r}$$

denote the number of times item i is classified in category j by the raters of the group and p_{ij} the corresponding proportion. We have

$$\sum_{j=1}^{K} p_{ij} = 1 \quad i = 1, \ldots, N.$$

Let y_{ij} denote the observed value of the random variable Y_{ij} corresponding to the category assignment of item i made by the isolated rater. Then, the data can be conveniently summarized in a two-way classification table (see Table 2) by defining the quantities

$$c_{jk} = \frac{1}{N} \sum_{i=1}^{N} p_{ij} y_{ik}, \quad j, k = 1, \ldots, K.$$

The *observed proportion of agreement* between the group of raters and the isolated rater is defined by

$$p_o = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} p_{ij} y_{ij} = \sum_{j=1}^{K} c_{jj}.$$

Table 2. Two-way classification table of the N items by the group of raters and the isolated rater.

	Isolated rater					
Group of raters	1	$\cdots$	j	$\cdots$	K	Total
1	c_{11}	$\ldots$	c_{1j}	$\ldots$	c_{1K}	$c_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	c_{j1}	$\ldots$	c_{jj}	$\ldots$	c_{jK}	$c_{j.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	c_{K1}	$\ldots$	c_{Kj}	$\ldots$	c_{KK}	$c_{K.}$
Total	$c_{.1}$	$\ldots$	$c_{.j}$	$\ldots$	$c_{.K}$	1

© 2009 The Authors. Journal compilation © 2009 VVS.

The marginal classification distribution of the isolated rater, namely,

$$y_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij}, \quad j = 1, \ldots, K$$

with $\Sigma_{j=1}^{K} y_j = 1$ and the marginal classification distribution of the group of raters,

$$p_j = \frac{1}{N}\sum_{i=1}^{N} p_{ij}, \quad j = 1, \ldots, K$$

with $\Sigma_{j=1}^{K} p_j = 1$ are needed to estimate the agreement expected by chance. The *proportion of agreement expected by chance* is given by

$$p_e = \sum_{j=1}^{K} p_j y_j = \sum_{j=1}^{K} c_{j.} c_{.j}$$

The degree of agreement κ between the population of raters and the isolated rater is then estimated by

$$\hat{\kappa} = \frac{p_o - p_e}{p_m - p_e}$$

where p_m corresponds to the maximum possible proportion of agreement derived from the data set. By extending the argument used for the binary case, it is easily seen that

$$p_m = \frac{1}{N}\sum_{i=1}^{N} \max_j p_{ij}. \tag{7}$$

Observe that in the calculation of p_m, no explicit use is made of category j in which the maximum occurs. Thus, in the case where the maximum is not unique, only the value of the maximum is actually important.

4.3 Ordinal scale ($K > 2$)

The estimation of the weighted agreement index is simply done by introducing weights in the estimations previously defined. Hence,

$$\hat{\kappa}_W = \frac{p_{o,w} - p_{e,w}}{p_{m,w} - p_{e,w}}$$

with

$$p_{o,w} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} p_{ij} y_{ik}$$

$$p_{e,w} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} p_j y_k$$

© 2009 The Authors. Journal compilation © 2009 VVS.

$$p_{m,w} = \frac{1}{N}\sum_{i=1}^{N} \max_{j}\left(\sum_{k=1}^{K} w_{jk} p_{ik}\right).$$

The unweighted agreement index $\hat{\kappa}$ can be obtained using the weights $w_{jj} = 1$ and $w_{jk} = 0, j \neq k$.

5 Asymptotic sampling variance

The Jackknife method (Efron and Tibshirani, 1993) can be used to determine the sampling variance of the agreement index. Suppose that the agreement between the isolated rater and the population of raters was estimated on a random sample of N items. Let $\hat{\kappa}_N$ denote the agreement index and $\hat{\kappa}_{N-1}^{(i)}$ the estimated agreement index when observation i is deleted. These quantities are used to determine the pseudo-values

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

The Jackknife estimator of the agreement index is then defined by

$$\tilde{\kappa}_N = \frac{1}{N}\sum_{i=1}^{N} \hat{\kappa}_{N,i}$$

with variance

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

The bias of the Jackknife estimator is estimated by

$$\text{Bias}(\tilde{\kappa}_N) = (N-1)\{\tilde{\kappa}_N - \hat{\kappa}_N\}$$

6 Consensus approach

As already mentioned, the consensus approach consists in summarizing the responses given by the raters of the group in a unique quantity for each item. Very often, the consensus category is taken as the modal category (majority rule) or the category chosen by a prespecified proportion of raters (e.g. $\geq 80\%$). A random variable Z_{ij} is then defined, equal to 1 if category j corresponds to the consensus category given by the population $\mathcal{R}$ of raters for item i, and equal 0 otherwise. Evidently, a consensus may not always be defined. For example, on a multinomial scale, one could have two modal categories or no category chosen by the prespecified proportion of raters. Therefore, suppose that on the N items drawn from population $\mathcal{I}$, a consensus can only be defined on $N_C \leq N$ items. Let $\mathcal{I}_C$ denote the sub-population of items on which a consensus is always possible. If z_{ij} denotes the observed value of the

© 2009 The Authors. Journal compilation © 2009 VVS.

random variable Z_{ij}, we have $\Sigma_{j=1}^{K} z_{ij} = 1$ and the agreement between the population of raters and the isolated raters is reduced to the case of two isolated raters. The Cohen or weighted kappa coefficient can then be estimated. Note that the strength of the consensus is not captured by the random variable Z_{ij}. For example, on a binary scale, using the majority rule, we have $Z_{ij} = 1$ regardless of the value of P_{ij} as long as $P_{ij} > 0.5$. It can easily be shown that our method and the consensus approach are equivalent only in two particular cases: first, when there is only one rater in the group of raters ($R = 1$) and secondly, when $\mathcal{I}_C = \mathcal{I}$ and there is perfect agreement in the population of raters (ICC = 1).

7 Examples

7.1 Syphilis serology

A proficiency testing programme for syphilis serology was conducted by the College of American Pathologists (CAP). For the fluorescent treponemal antibody absorption test (FTA-ABS), three reference laboratories were identified and considered as experts in the use of that test. During 1974, 40 syphilis serology specimens were tested independently by the three reference laboratories. Williams (1976) presented results obtained by the three reference laboratories and an additional participant (noted L) for 28 specimens (see Table 3). Each specimen was classified as non-reactive (NR), borderline (BL) or reactive (RE). Note that discordances occurred between the three reference laboratories for seven specimens. Data are also summarized in a two-way classification table (Table 4) as explained in Section 4.2. In this example $R = 3$, $K = 3$ and $N = 28$.

Using the quadratic weighting scheme, the weighted coefficient of agreement $\hat{\kappa}_W$ ($\pm$SE) between the participant and the three reference laboratories, as defined in Section 3, was equal to 0.79 ($\pm$0.06). When applying the consensus approach based on the majority rule, we found a weighted kappa coefficient of 0.76 ($\pm$0.06), but two specimens were eliminated because no consensus could be reached between the three reference laboratories. The weighted agreement index developed by Schouten (1982) amounted 0.73 ($\pm$0.07), while the ICC in the reference laboratory group was 0.68 ($\pm$0.06). Because of the lack of perfect agreement among the reference laboratories (ICC < 1), Schouten's agreement index can never be equal to 1 so that perfect agreement can never be attained. According to Equation 7, the non-weighted maximum attainable proportion was $p_m = 0.893$, while the corresponding value for the quadratic weighting scheme was $p_{m,w} = 0.973$. To derive the highest possible value of the proposed agreement index, consider the hypothetical laboratory H whose responses are given in Table 3. For this particular laboratory, as each specimen's result corresponds to the most frequent response given by the reference laboratories, our agreement index yields the perfect value of 1 ($\pm$0), while Schouten's index is only equal to 0.94 ($\pm$0.025). For the consensus approach, the kappa coefficient derived was

© 2009 The Authors. Journal compilation © 2009 VVS.

Table 3. Classification of 28 specimens for syphilis serology on a three-category scale (NR = non-reactive, BL = bordeline, RE = reactive) by two individual laboratories (L and H) and three reference laboratories (data from WILLIAMS, 1976).

	Participant		Reference		
Specimen	L	H^*	R1	R2	R3
1	RE	RE	RE	RE	RE
2	RE	RE	RE	RE	RE
3	BL	NR	NR	NR	NR
4	BL	NR	NR	NR	NR
5	BL	NR	NR	NR	NR
6	RE	RE	RE	RE	RE
7	BL	NR	NR	NR	NR
8	RE	RE	RE	RE	RE
9	NR	NR	NR	NR	NR
10	NR	NR	NR	NR	NR
11	RE	RE	RE	RE	RE
12	RE	BL	RE	BL	BL
13	RE	RE	RE	RE	RE
14	RE	BL	RE	BL	BL
15	RE	RE	RE	RE	RE
16	RE	BL	RE	NR	BL
17	RE	BL	RE	NR	BL
18	RE	RE	RE	RE	RE
19	RE	RE	RE	RE	RE
20	BL	NR	BL	NR	NR
21	RE	RE	RE	RE	RE
22	BL	NR	NR	NR	NR
23	BL	NR	BL	NR	NR
24	BL	NR	BL	NR	NR
25	RE	RE	RE	RE	RE
26	NR	NR	NR	NR	NR
27	RE	RE	RE	RE	RE
28	NR	NR	NR	NR	NR

*Hypothetical participant (see text).

also equal to 1, although two specimens (16 and 17) have to be excluded. Finally, it should be remarked that if the hypothetical laboratory H had supplied results different from BL for specimens 16 and 17, the non-weighted agreement coefficient obtained would still be 1 but the weighted version would yield a value less than 1 because of the weighting scheme ($\kappa_W = 0.958$).

Table 4. Two-way classification table of the 28 syphilis serology specimens as NR (non-reactive), BL (borderline) and RE (reactive) by three reference laboratories and participant L.

	Participant L			
Reference laboratories	NR	BL	RE	Total
NR	0.143	0.250	0.024	0.417
BL	0	0.036	0.071	0.107
RE	0	0	0.476	0.476
Total	0.143	0.286	0.571	1

© 2009 The Authors. Journal compilation © 2009 VVS.

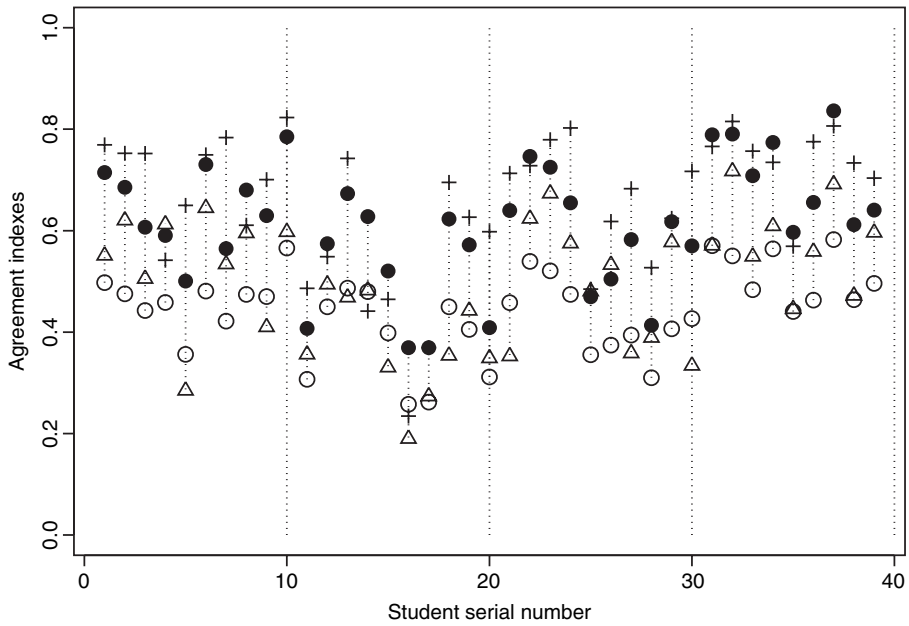

Fig. 1. Values of k_W (•), weighted κ coefficients using the majority (△) and the 50% (+) rules and weigthed agreement index of Schouten (○) for the 39 students passing the SCT.

7.2 Script Test of Concordance

The Script Test of Concordance (SCT) is used in medicine to evaluate the ability of physicians or medical students (isolated raters) to solve clinical situations not clearly defined (Charlin *et al.*, 2002). The complete test consists of a number of items $(1,\ldots,N)$ to be evaluated on a five-point Likert scale ($K=5$). Each item represents a clinical situation likely to be seen in real-life practice and a potential assumption is proposed with it. The situation has to be unclear, even for an expert. The task of the student or the physician being evaluated is to consider the effect of additional evidence on the suggested assumption. In this respect, the candidate has 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 basically the only possible one. The questionnaire is also given to a panel of experts (raters $1,\ldots,R$). The problem is to evaluate the agreement between each individual medical student and the panel of experts.

Between 2003 and 2005, an SCT was proposed to students training in general practice (Vanbelle *et al.*, 2007). The SCT consisted of 34 items relating possible situations encountered in general practice. There were 39 students passing the test and completing the entire questionnaire. Their responses were confronted to the responses of a panel of 11 experts. The intraclass correlation coefficient in the group of experts was 0.22 (± 0.04). The individual $\hat{\kappa}_W$ coefficients for the 39 students were

© 2009 The Authors. Journal compilation © 2009 VVS.

computed using the quadratic weighting scheme. Values ranged between 0.37 and 0.84 and the mean agreement index ± standard deviation (SD) was 0.61 ± 0.12. Schouten's weighted index scores averaged 0.44 ± 0.08 (range: 0.26–0.58).

Using the consensus method, where consensus was defined as either the majority of the raters or a proportion of at least 50% of the raters, 2 (6%) and 12 (35%) items had to be omitted from the analysis, respectively, because no consensus was reached among the experts. The mean weighted kappa values for the 39 students was equal to 0.49 ± 0.13 (range: 0.19–0.72) with the majority rule and 0.66 ± 0.14 (range: 0.23–0.82) with the 50% rule. Figure 1 displays the individual agreement coefficients relative to each student for the various methods. Marked differences can be seen on the graph depending on the approach used. A ranking of the students was needed for selection purposes. The ranking changed notably for some students according to the agreement index calculated. For example, student no. 39 ranked at the 16th place with the new approach, the 9th place with Schouten's index, the 10th place using the majority rule and at 20th place using the 50% rule.

8 Discussion

The method described in this paper was developed to quantify the agreement between an isolated rater and a group of raters judging items on a categorical scale. The group of raters is seen as a well-defined entity, a reference or gold-standard group with its own heterogeneity, whereas the isolated rater comes from a distinct population. Therefore, the marginal classification probabilities of the isolated rater and of the population of raters were basically assumed to be different ($\pi \neq \pi^*$). In the SCT example, it is realistic to admit that each student differs from the group of experts by the knowledge he/she acquired so far in clinical decision-making. Although the group of raters was seen as the 'reference' group in the present study, the theory is equally applicable to the case where the isolated rater represents the expert, at least as long as a single agreement index is looked for between them. When neither the isolated rater nor the group of raters is considered as the gold standard, an intraclass version of the proposed agreement index can be derived. The latter reduces to the ICC (Kraemer, 1979) in case of two isolated raters, by assuming that the isolated rater and the group of raters come from the same population ($\pi = \pi^*$). The agreement index was conveniently developed on a population-based model, allowing an easy extension from dichotomous to multinomial scales and the use of weighted agreement coefficients. It also leads to a less restrictive definition of perfect agreement. Indeed, the isolated rater and the group of raters were defined to be in 'perfect agreement' when their respective classifications of items were linearly related and equal on average, without perfect agreement among all raters in the group (ICC < 1). It was shown that under this assumption and the additional assumption of perfect agreement within the population of raters (ICC = 1), the proposed agreement index κ is algebraically equivalent to the agreement coefficient derived by Schouten (1982). In other terms, the present approach is based on less

© 2009 The Authors. Journal compilation © 2009 VVS.

stringent assumptions than those made by Schouten. This was illustrated on the syphilis example where it was not possible for Schouten's agreement index to achieve the maximum value of 1, contrary to the new agreement index. The latter further overcomes the shortcomings of the widely used consensus method, in particular the fact that a decision is not required for items lacking a consensus in the group. It should be remarked, however, that for items lacking consensus among the members of the group, the responses given by the isolated rater can lead to different kappa values depending on the scheme used (weighted or non-weighted) as demonstrated by the hypothetical laboratory in Williams' example. The new agreement index also takes into account the existing heterogeneity in the group of raters while the strength of consensus, as already indicated, is completely ignored in the consensus method. Finally, as illustrated in the SCT example and pointed out by Salerno *et al.* (2003) and Miller *et al.* (2004), the results may vary markedly according to the definition of the consensus method used.

The notion of perfect agreement appears to play a major role in the definition of the new agreement coefficient and particularly of its maximum value of 1. Here, the population of raters is seen as a whole, a single entity composed of equally valued members but displaying heterogeneity in their judgements of items. Hence, perfect agreement is defined between the isolated rater and the population itself, not between the isolated rater and the individual members of the population. As a consequence, agreement may be perfect without forcing all raters, including the isolated one, to classify all items in the same way. The present definition also does not preclude that the agreement between the isolated rater and the population may be better than the agreement between the population and some of its individual members. In other terms, the isolated rater can perform better than some of the experts. This may sound somewhat contradictory in the context of a gold standard. In Schouten's view, an agreement value of 1 can only be achieved when all raters of the population and the isolated rater perfectly and thoroughly agree in allocating items. A gold standard generally represents some practically not attainable but only approachable level or quantity determined by a single reference method. There are situations, however, where a gold standard may result from the application of several reference methods or the opinions of several experts, without necessarily achieving a perfect consensus on all items. In a medical context, the various responses of an expert group may not only reflect the absence of a clear consensus among experienced physicians but also the fuzzy character of the clinical situation at hand. As seen with Williams' syphilis serology data, major discrepancies were observed in the responses given by the three reference laboratories for some of the assayed specimens. It is therefore our opinion that proficiency testing programmes should allow for the fact that a particular non-reference laboratory is in perfect agreement with the references laboratories without being in perfect agreement with each of them separately, unlike Schouten's index.

While in theory we may assume that there is always a category of the K-categorical scale with a maximum proportion of raters for each item, it is not necessarily the

© 2009 The Authors. Journal compilation © 2009 VVS.

case in practice. There may indeed be a maximum shared by two or more categories, which have to be compared with the category chosen by the isolated rater for this item (see hypothetical laboratory example in Table 3). However, as mentioned previously, this has virtually no impact on the agreement coefficient obtained. In other terms, two distinct isolated raters will yield the same agreement coefficient (ignoring the weighting scheme) although their response profile is not exactly identical.

In sum, the proposed kappa coefficient provides a useful alternative to the consensus method and to Light's approach. It also generalizes the agreement index proposed by Schouten (1982) as well as Cohen's kappa coefficient while keeping its attractive properties.

Acknowledgment

The authors are much indebted to one of the reviewers whose constructive remarks considerably improved the content of the paper.

9 Appendix

9.1 Perfect agreement when $K=2$

Equivalence 1. *The definition of perfect agreement,* $E(P_i)=E(Y_i)=\pi^{**}$ *and* $\text{corr}(P_i, Y_i)=1$*, is equivalent to writing* $P_i=\pi^{**}(1-\sqrt{ICC})+\sqrt{ICC}\,Y_i$.

Proof. Indeed, $\rho=1$ leads to the linear relation $P_i=a+bY_i$. This implies

$$E(P_i)=\pi^{**}=E(a+bY_i)=a+b\pi^{**}$$
$$\text{var}(P_i)=\sigma^2=\text{var}(a+bY_i)=b^2\ \text{var}(Y_i)=b^2\pi^{**}(1-\pi^{**})$$

Thus, $a=(1-b)\pi^{**}$ and $P_i=(1-b)\pi^{**}+bY_i$.

As

$$\text{ICC}=\frac{\sigma^2}{\pi(1-\pi)}=b^2\frac{\pi^{**}(1-\pi^{**})}{\pi^{**}(1-\pi^{**})}=b^2,$$

we have

$$P_i=\pi^{**}(1-\sqrt{\text{ICC}})+\sqrt{\text{ICC}}\,Y_i. \qquad \square$$

9.2 Perfect agreement for $K>2$

Equivalence 2. *If* Π_M *is defined by*

$$\Pi_M=\sum_{j=1}^{K}E[(\pi_j^{**}+(1-\pi_j^{**})\sqrt{ICC_j})\,Y_{ij}]$$

© 2009 The Authors. Journal compilation © 2009 VVS.

where $E(P_{ij}) = E(Y_{ij}) = \pi_j^{**}$ *and* ICC_j *denotes the intraclass kappa coefficient relative to category* j $(j = 1, \ldots, K)$ *in the population of raters, we have*

$$\sum_{j=1}^{K} \Pi_{[j]M} = 2\Pi_M + K - 2$$

where $\Pi_{[j]M}$ *corresponds to the quantity described in the binary case* $(K = 2)$ *when the nominal scale is dichotomized by grouping all categories other than category* j *together.*

Proof. When the population of raters and the isolated rater are in perfect agreement, we have from Equivalence 1

$$P_{ij} = \pi_j^{**}(1 - \sqrt{\mathrm{ICC}_j}) + \sqrt{\mathrm{ICC}_j}\, Y_{ij}$$

Therefore,

$$\begin{aligned}
\Pi_M = E\left[\sum_{j=1}^{K} P_{ij}\, Y_{ij}\right] &= E\left[\sum_{j=1}^{K} \left(\pi_j^{**} + \left(1 - \pi_j^{**}\right)\sqrt{\mathrm{ICC}_j}\, Y_{ij}\right) Y_{ij}\right] \\
&= \sum_{j=1}^{K} (\pi_j^{**} + (1 - \pi_j^{**})\sqrt{\mathrm{ICC}_j})\pi_j^{**} \\
&= \sum_{j=1}^{K} \left(\pi_j^{**2} + \sigma_j^{**2}\frac{\sigma_j}{\sigma_j^{**}}\right) \\
&= \sum_{j=1}^{K} \left(\pi_j^{**2} + \sigma_j \sigma_j^{**}\right)
\end{aligned}$$

From Equation 3,

$$\begin{aligned}
\sum_{j=1}^{K} \Pi_{[j]M} &= \sum_{j=1}^{K} \left(1 - 2\pi_j^{**}(1 - \pi_j^{**})\left(1 - \sqrt{\mathrm{ICC}_j}\right)\right) \\
&= \sum_{j=1}^{K} \left(1 - 2\sigma_j^{**2}\frac{\sigma_j^{**} - \sigma_j}{\sigma_j^{**}}\right) \\
&= \sum_{j=1}^{K} 1 - 2\sum_{j=1}^{K} \sigma_j^{**2} + 2\sum_{j=1}^{K} \sigma_j^{**}\sigma_j \\
&= K - 2 + 2\sum_{j=1}^{K} \left(\pi_j^{**2} + \sigma_j \sigma_j^{**}\right) \\
&= 2\Pi_M + K - 2
\end{aligned}$$

□

© 2009 The Authors. Journal compilation © 2009 VVS.

References

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

Cicchetti, D. V. and T. Allison (1971), A new procedure for assessing reliability of scoring EEG sleep recordings, *American Journal of EEG Technology* **11**, 101–109.

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 or partial credit, *Psychological Bulletin* **70**, 213–220.

Eckstein, M. P., T. D. Wickens, G. Aharonov, G. Ruan, C. A. Morioka and J. S. Whiting (1998), Quantifying the limitations of the use of consensus expert commitees in ROC studies, *Proceedings SPIE: Medical Imaging 1998: Image Perception* **3340**, 128–134.

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

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

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.

Kalant, N., M. Berlinguet, J. G. Diodati, L. Dragatakis and F. Marcotte (2000), How valid are utilization review tools in assessing appropriate use of acute care beds? *Canadian Medical Association Journal* **162,** 1809–1813.

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

Kraemer, H. C., V. S. Periyakoil and A. Noda (2002), Kappa coefficient in medical research *Tutorials in Biostatistics* **1**, 85–105.

Landis, J. R. and G. G. Koch (1977), An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, *Biometrics* **33**, 363–374.

Light, R. J. (1971), Measures of response agreement for qualitative data: some generalizations and alternatives, *Psychological Bulletin* **76**, 365–377.

Miller, D. P., K. F. O'Shaughnessy, S. A. Wood and R. A. Castellino (2004), Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions, *Proceedings SPIE: Medical Imaging 1998: Image Perception, Observer Performance and Technology Assessment* **5372**, 173–184.

Ruperto, N., A. Ravelli, S. Oliveira, M. Alessio, D. Mihaylova, S. Pasic, E. Cortis, M. Apaz, R. Burgos-Vargas, F. Kanakoudi-Tsakalidou, X. Norambuena, F. Corona, V. Gerloni, S. Hagelberg, A. Aggarwal, P. Dolezalova, C. M. Saad, S. C. Bae, R. Vesely, T. Avcin, H. Foster, C. Duarte, T. Herlin, G. Horneff, L. Lepore, M. van Rossum, L. Trail, A. Pistorio, B. Andersson-Gare, E. H. Giannini and A. Martini (2006), Pediatric Rheumatology International Trials Organization. The Pediatric Rheumatology International Trials Organization/American College of Response to Therapy in Juvenile Systemic Lupus Erythematosus: prospective validation of the definition of improvement, *Arthritis and Rheumatism (Arthritis Care and Research)* **55**, 355–363.

Salerno, S. M., P. C. Alguire and S. W. Waxman (2003), Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence, *Annals of Internal Medicine* **138**, 751–760.

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

Smith, R., A. J. Copas, M. Prince, B. George, A. S. Walker and S. T. Sadiq (2003), Poor sensitivity and consistency of microscopy in the diagnosis of low grade non-gonococcal urethritis, *Sexually Transmitted Infections* **79**, 487–490.

Soeken, K. L. and P. A. Prescott (1986), Issues in the use of kappa to estimate reliability, *Medical care* **24**, 733–741.

© 2009 The Authors. Journal compilation © 2009 VVS.

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

Williams, G. W. (1976), Comparing the joint agreement of several raters with another rater, *Biometrics* **32**, 619–627.

Received: December 2007. Received: November 2008.

© 2009 The Authors. Journal compilation © 2009 VVS.