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

Correspondence should be sent to

E-Mail: sophie.vanbelle@ulg.ac.be

Phone: +3243662590

Fax: +3243662596

Abstract

The agreement between two raters judging items on a categorical scale is traditionally measured 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 coefficient, which is defined on a population-based model, requires a specific definition of the concept of perfect agreement but possesses the same properties as Cohen's kappa coefficient. Further, it reduces to the classical kappa when there is only one rater in the group. An intraclass and a weighted versions of the coefficient are also introduced. The new approach overcomes the problem of consensus and generalizes Schouten's index. The sampling variability of the agreement coefficient is derived by the Jackknife technique. The method is illustrated on published syphilis data and on data collected from a study assessing the ability of medical students in diagnostic reasoning.

Keywords: kappa coefficient; nominal scale; ordinal scale.

1 INTRODUCTION

Cohen (1960) introduced the kappa coefficient $\kappa = (p_o - p_e)/(1 - p_e)$ to quantify the agreement between two raters classifying N items on a binary or nominal 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 a value 1 when the agreement between the two raters is perfect and 0 when the observed agreement is equal to the agreement expected by chance. There are situations where the agreement between an isolated rater and a group of raters is needed. For example, each of a series of individuals may be assessed against a group of experts and a ranking of the individuals may be required. Conversely, agreement may be searched between a group of users and a gold standard. Usually in such instances, a consensus is determined in the group of raters and the problem is reduced to the case of measuring agreement between the isolated rater and the consensus in the group (Landis and Koch 1977, Soeken and Prescott 1986, Salerno *et al.* 2003). The consensus may be defined as the category chosen by a given proportion of the raters in the group (for example, Ruperto *et al.* (2006) defined the 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, the problem of how to handle items without consensus arises. Ruperto *et al.* (2006) discarded all patients 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) investigated 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) examined the possibility to obtain different conclusions by using different rules of consensus. Light (1971) developed a statistic for comparing the joint agreement of several raters with a gold standard. This statistic is a mixture of the proportions of concordant pairs obtained between each of the rater in the group and the gold standard (the isolated rater). His method leads to tedious calculations, does not quantify the agreement between the gold standard and the group of raters and the calculations have not been extended to the case of a group including more than 3 raters. 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. Indeed, he compared the mean proportion of concordant items between the isolated rater and each rater in the group to the mean proportion of concordant items between all possible pairs of raters among the group of raters. The ratio derived (Williams' index) is compared to the value of 1. Unfortunately, the coefficient proposed by Williams (1976) does not correct for agreements due to chance and does not quantify the agreement between the isolated rater and the group of raters. Finally, Schouten (1982) developed a method of hierarchical clustering based on pairwise weighted agreement measures to select one or more homogeneous subgroups of raters when several raters classify items on a nominal or an ordinal scale. Hereafter, we propose a coefficient for quantifying the agreement between an isolated rater and a group of raters, which overcomes the problem of consensus, generalizes the approach of Schouten (1982) and possesses the same properties as Cohen's kappa coefficient.

2 DEFINITION OF THE AGREEMENT INDEX

2.1 Binary scale (K=2)

Consider a population $\mathcal{I}$ of items and a population $\mathcal{R}$ of raters. Suppose that the items have to be classified on a binary scale by the population of raters and by an independent isolated rater. Let $X_{i,r}$ be the random variable such that $X_{i,r} = 1$ if a randomly selected rater r of the population $\mathcal{R}$ classifies a randomly selected item i of population $\mathcal{I}$ in category 1 and $X_{i,r} = 0$ otherwise. Let $E(X_{i,r}) = P(X_{i,r} = 1) = p_i$ over the population of raters. Then, over the population of items, let $E(p_i) = \pi$ and $\sigma^2 = var(p_i)$. In the same way, 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 $var(Y_i) = \sigma^{*2} = \pi^*(1-\pi^*)$. Finally, let ICC denote the intraclass correlation coefficient in the population of raters

Table 1: Theoretical model for the classification of a randomly selected item i on a binary scale by the population of raters $\mathcal{R}$ and the isolated rater

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

(Fleiss 1981)

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

and ρ the correlation between p_i and Y_i over $\mathcal{I}$

$$\rho = \frac{E(p_iY_i)-\pi\pi^*}{\sigma\sigma^*}. \tag{2}$$

Using these definitions, a 2×2 table can be constructed cross-classifying the population of raters $\mathcal{R}$ and the isolated rater with respect to the binary scale (Table 1).

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

$$\Pi_i = p_iY_i + (1-p_i)(1-Y_i) \tag{3}$$

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

$$\Pi_T = E(\Pi_i) = \pi\pi^* + (1-\pi)(1-\pi^*) + 2\rho\sigma\sigma^* \tag{4}$$

By definition, the population of raters and the isolated rater are considered to be in "perfect agreement" if and only if

$$\pi = \pi^* = \pi^{**} \text{ and } \rho = 1. \tag{5}$$

In terms of the random variables p_i and Y_i over $\mathcal{I}$ this is equivalent to writing

$$p_i = \pi^{**}(1 - \sqrt{ICC}) + \sqrt{ICC}Y_i \tag{6}$$

It follows from Equation 4 that the maximum attainable probability of perfect agreement is given by

$$\Pi_M = 1 - 2\pi^{**}(1-\pi^{**})(1-\sqrt{ICC}) \tag{7}$$

which turns out to be equal to 1 only if $ICC = 1$, i.e. that there is perfect agreement between all raters in population $\mathcal{R}$, or trivially if $\pi^{**} = 0$ or 1.

Then, the coefficient of agreement between the population of raters and the isolated rater is defined in a kappa-like way:

$$\kappa = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E} \tag{8}$$

where Π_E is the agreement expected by chance, i.e., the probability that the population of raters and the isolated rater agree under the independence assumption

$(E(p_iY_i) = E(p_i)E(Y_i))$, defined by

$$\Pi_E = \pi\pi^* + (1-\pi)(1-\pi^*) \tag{9}$$

Note that $\Pi_T = \Pi_E$ when there is no correlation between the ratings of the population of raters and the isolated rater ($\rho = 0$) or when there is no variability in the classification made by the populations of raters ($\sigma^2 = 0$) or by the isolated rater ($\sigma^{*2} = 0$).

An intraclass version of the agreement index κ_I may be derived by assuming that $\pi = \pi^* = \pi^{**}$. It leads to

$$\kappa_I = \rho = \frac{E(p_iY_i) - \pi^{**}}{\sigma\sqrt{\pi^{**}(1-\pi^{**})}} \tag{10}$$

2.2 Multinomial scale (K>2)

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

$$\kappa = \frac{\sum_{j=1}^{K}(\Pi_{[j]T} - \Pi_{[j],E})}{\sum_{j=1}^{K}(\Pi_{[j]M} - \Pi_{[j]E})} = \frac{\Pi_T - \Pi_E}{\Pi_M - \Pi_E} \tag{11}$$

where $\Pi_{[j]T}$, $\Pi_{[j]E}$ and $\Pi_{[j]M}$ correspond to the quantities described in the 2×2 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 respectively by

$$\Pi_T = \sum_{j=1}^{K} E(p_{ij}Y_{ij}); \quad \Pi_E = \sum_{j=1}^{K} \pi_j\pi_j^*;$$

$$\Pi_M = \sum_{j=1}^{K} E((\pi_j^{**} + (1 - \pi_j^{**})\sqrt{ICC_j})Y_{ij}) \tag{12}$$

where p_{ij} denotes the probability for a randomly selected item i to be classified in category j ($j = 1, \cdots, K$) by the population of raters with $E(p_{ij}) = \pi_j$ and 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 intraclass correlation coefficient relative to category j ($j = 1, \cdots, K$) in the population of raters.

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 expected by chance ($\Pi_T < \Pi_E$).

2.3 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),

$$\kappa_W = \frac{\Pi_{T,W} - \Pi_{E,W}}{\Pi_{M,W} - \Pi_{E,W}} \tag{13}$$

with

$$\Pi_{T,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E(p_{ij} Y_{ik}); \tag{14}$$

$$\Pi_{E,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} \pi_j \pi_k^*; \tag{15}$$

$$\Pi_{M,W} = \sum_{j=1}^{K}\sum_{k=1}^{K} w_{jk} E(\pi^{**}(1-\sqrt{ICC_j})Y_{ik}) + \sum_{k=1}^{K}\sqrt{ICC_k}. \qquad (16)$$

In general, $0 \leq w_{jk} \leq 1$ and $w_{kk} = 1$, $(j,k = 1,\cdots,K)$. Cicchetti and Allison (1971) have defined absolute weights $w_{jk} = 1 - \dfrac{|j-k|}{K-1}$ whereas Fleiss and Cohen (1973) suggested quadratic weights $w_{jk} = 1 - \left(\dfrac{j-k}{K-1}\right)^2$.

3 ESTIMATION OF THE PARAMETERS

Suppose that a random sample of N items drawn from population $\mathcal{I}$ is classified on a K-categorical scale by a random sample (group) of size R from the population of raters $\mathcal{R}$ and by an independent isolated rater.

3.1 Binary scale ($K = 2$)

Let $x_{i,r}$ denotes the observed value of the random variable $X_{i,r}$ denoting the rating of rater r of the population $\mathcal{R}$ $(i = 1,\cdots,N; r = 1,\cdots,R)$. Let y_i denotes the observed value of the random variable Y_i representing the rating of the isolated rater. Then, let $n_i = \sum_{r=1}^{R} x_{i,r}$ denotes the number of times the item i is classified in category 1 by the group of raters and let $\widehat{p}_i = n_i/R$ denote the corresponding proportions $(i = 1,\cdots,N)$.

The intraclass correlation coefficient in the group of raters is estimated by

(Fleiss 1981)

$$\widehat{ICC} = 1 - \frac{\sum_{i=1}^{N} n_i(R - n_i)}{RN(N-1)p(1-p)} \tag{17}$$

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

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

The probability that the population of raters and the isolated rater agree is estimated by the *observed proportion of agreement*,

$$\widehat{\Pi}_T = p_o = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p}_i y_i + (1-\widehat{p}_i)(1-y_i)). \tag{18}$$

Clearly, $p_o = 1$ if the raters of the group and the isolated rater classify each item in the same category and $p_o = 0$ if the isolated rater systematically classifies items in a category never chosen by the group of raters.

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

$$p_e = py + (1-p)(1-y) \tag{19}$$

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

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e} \tag{20}$$

where p_m corresponds to the maximum possible proportion of agreement derived by the data. We have

$$p_m = \frac{1}{N}\sum_{i=1}^{N} max(\widehat{p}_i, 1 - \widehat{p}_i). \tag{21}$$

3.2 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 $(r = 1, \cdots, R)$ of the group classified item i $(i = 1, \cdots, N)$ in category j $(j = 1, \cdots, K)$ and equal to 0 otherwise. In the same way, let y_{ij} denote the observed value of the random variable Y_{ij} corresponding to the rating of the isolated rater. Let $n_{ij} = \sum_{r=1}^{R} x_{ij,r}$ denotes the number of times the item i is classified in category j by the raters of the group and let $\widehat{p}_{ij}$ denote the corresponding proportions. We have $\sum_{j=1}^{K} \widehat{p}_{ij} = 1$, $(i = 1, \cdots, N)$. The data can be conveniently summarized in a 2-way classification table (see Table 2) by defining the quantities

$$c_{jk} = \frac{1}{N}\sum_{i=1}^{N} \widehat{p}_{ij} y_{ik},\ \ j, k = 1, \cdots, K \tag{22}$$

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

$$p_o = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij} y_{ij} = \sum_{j=1}^{K} c_{jj} \tag{23}$$

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

The marginal classification distribution of the isolated rater, namely,

$$y_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij}, \quad j = 1, \cdots, K \tag{24}$$

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

$$p_j = \frac{1}{N}\sum_{i=1}^{N} \widehat{p}_{ij}, \ j = 1 \cdots, K \tag{25}$$

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

$$p_e = \sum_{j=1}^{K} p_j y_j = \sum_{j=1}^{K} c_{j.} c_{.j} \tag{26}$$

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

$$\widehat{\kappa} = \frac{p_o - p_e}{p_m - p_e} \quad (27)$$

where p_m corresponds to the maximum possible proportion of agreement derived from the data,

$$p_m = \frac{1}{N} \sum_{i=1}^{N} max_j p_{ij}. \quad (28)$$

Note that when $R = 1$, $p_m = 1$ and the agreement coefficient $\widehat{\kappa}$ reduces to the classical Cohen's kappa coefficient defined in the case of two isolated raters.

The intraclass correlation coefficient in the group of raters is estimated by (Fleiss 1981)

$$\widehat{ICC} = 1 - \frac{NR^2 - \sum_{i=1}^{N} \sum_{j=1}^{K} n_{ij}^2}{NR(R-1) \sum_{j=1}^{K} p_j(1-p_j)} \quad (29)$$

3.3 Ordinal scale ($K > 2$)

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

$$\widehat{\kappa}_W = \frac{p_{o,w} - p_{e,w}}{p_{m,w} - p_{e,w}} \quad (30)$$

with

$$\begin{aligned} p_{o,w} &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_{ij} y_{ik} \\ p_{e,w} &= \sum_{j=1}^{K} \sum_{k=1}^{K} w_{jk} p_j y_k \\ p_{m,w} &= \frac{1}{N} \sum_{i=1}^{N} max_j (\sum_{k=1}^{K} w_{jk} p_{ik}). \end{aligned} \quad (31)$$

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

4 ASYMPTOTIC SAMPLING VARIANCE

The Jackknife method (Efron and Tibshirani, 1993) was 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 $\widehat{\kappa}_N$ denote that agreement index and $\widehat{\kappa}_{N-1}^{(i)}$ denote the estimated agreement index when observation i is deleted. These quantities are used to determine the pseudo-values

$$\widehat{\kappa}_{N,i} = N\widehat{\kappa}_N - (N-1)\widehat{\kappa}_{N-1}^{(i)} \tag{32}$$

The Jackknife estimator of the agreement index is then defined by

$$\widetilde{\kappa}_N = \frac{1}{N}\sum_{i=1}^{N} \widehat{\kappa}_{N,i} \tag{33}$$

with variance

$$var(\widetilde{\kappa}_N) = \frac{1}{N}\left\{\frac{1}{N-1}\sum_{i=1}^{N}(\widehat{\kappa}_{N,i} - \widehat{\kappa}_N)^2\right\} \tag{34}$$

The bias of the Jackknife estimator is estimated by

$$Bias(\widetilde{\kappa}_N) = (N-1)\left\{\widetilde{\kappa}_N - \widehat{\kappa}_N\right\} \tag{35}$$

5 CONSENSUS APPROACH

The consensus approach consists in summarizing the responses given by the raters of the group in a unique quantity. Most approaches define the modal category (majority rule) or the category chosen by a prespecified proportion of raters ($\geq 50\%$) as the consensus category. A random variable Z_{ij} is then defined to be equal to 1 if category j corresponds to the consensus category given by the population $\mathcal{R}$ of raters for item i and equal to 0 otherwise. It is obvious that a consensus may not always be defined. For example, on a multinomial scale, we 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 random variable Z_{ij}, we have $\sum_{j=1}^{K} z_{ij} = 1$ and the agreement between the population of raters and the isolated raters is reduced to the case of 2 isolated raters. The Cohen intraclass or weighted kappa coefficient can then be estimated. Note that the strenght of the consensus is not taken into account by the random variable Z_{ij}. For example on a binary scale, using the majority rule, we will have $Z_{ij} = 1$ if $p_{ij} = 0.6$ but also if $p_{ij} = 0.9$. It can easily be shown that the new methodology defined and the consensus approach are equivalent only in two particular cases, firstly 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

$(ICC = 1)$.

6 EXAMPLES

6.1 Syphilis serology

A proficiency testing program for syphilis serology was conducted by the College of American Pathologists (CAP). For the fluorescent treponemal antibody absorption test (FTA-ABS), 3 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 3 reference laboratories. Williams (1976) presented data for 28 specimens. To evaluate the performance of a participant, the agreement between the participant and the 3 reference laboratories had to be evaluated. The data are summarized in a two-way classification table (Table 3) as explained is section 2.3.

Using the quadratic weighting scheme, the weighted coefficient of agreement $\widehat{\kappa}_W$ amounted 0.79 ± 0.06. When applying the consensus approach based on the majority rule, we found a weighted kappa coefficient of 0.76 ± 0.06. Remark that 2 specimens were eliminated because no consensus was found in the group of the 3 reference laboratories. Finally, the weighted agreement measure developed by Schouten (1982) was 0.73 ± 0.07. Note that the intraclass correlation coefficient was 0.68 ± 0.06 in the group of raters.

Table 3: Two-way classification table of the 28 syphilis serology specimens as NR (non-reactive), B (borderline) and R (reactive) by 3 reference laboratories and a participant

	Participant			
Reference laboratories	NR	B	R	Total
NR	0.143	0.250	0.024	0.417
B	0	0.036	0.071	0.107
R	0	0	0.476	0.476
Total	0.143	0.286	0.571	1

6.2 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 $(1, \cdots, N)$ to be evaluated on a 5-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 and (+2) The assumption is practically the only possibility. The questionnaire is also given to a panel of experts (raters $1, \cdots, R$). The problem is to evaluate the agreement between each individual medical student and the panel of experts.

Between 2003 and 2005, the SCT was proposed to students specializing in general practice at the University of Liège, Belgium (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 was 0.22 ± 0.04 in the group of experts. Using the quadratic weighting scheme, the individual $\widehat{\kappa}_W$ coefficients for the 39 students ranged between 0.37 and 0.84. The mean value ($\pm SD$) was 0.61 ± 0.12.

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, respectively 2 (6%) and 12 (35%) items had to be omitted from the analysis because no consensus was reached among the raters. The mean weighted kappa values for the 39 students were equal to 0.49 ± 0.13 (range: 0.19-0.72) and 0.66 ± 0.14 (range: 0.23-0.82) with the majority and the 50% rules, respectively. Figure 1 displays the individual agreement coefficients relative to each student for the different methods. A ranking of the student was needed in order to select only the best students. The

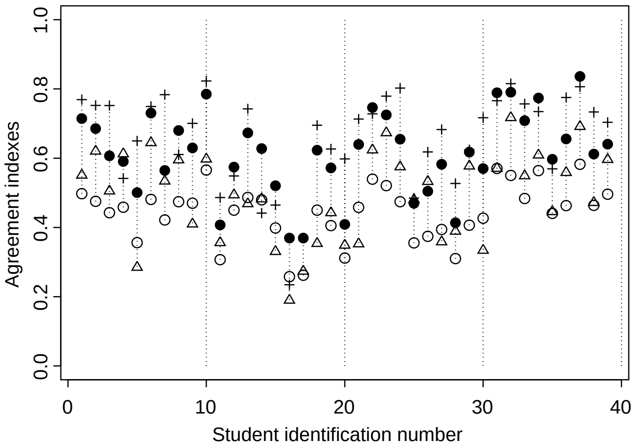

Figure 1: Values of κ_W ($\bullet$), weighted κ coefficients using the majority ($\triangle$) and the 50% (+) rules and weigthed agreement index of Schouten ($\circ$) for the 39 students passing the SCT

ranking changed markedly for some students according to the method used. For example, student No. 39 ranked at the 16th place with the new approach, the 9th place with Schouten index, the 10th place using the majority rule and at 20th place using the 50% rule.

7 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. A population-based approach was used but in case of a fixed group of raters, estimates are replaced by actual values. The derived agreement index κ possesses the same properties as Cohen's kappa coefficient (Cohen 1960) and reduces to it if there is only one rater in the group. The isolated rater and the group of raters are defined to be in perfect agreement when they have the same probability, for each item, to classify this item in a given category and the correlation coefficient between the isolated rater and the population of raters is equal to 1. It can be shown that with the additional assumption of perfect agreement in 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, perfect agreement can be reached between the isolated rater and the population of raters even if no perfect agreement occurs in the population of raters unlike the agreement index of Schouten (1982). The new approach is equivalent to the consensus approach when it is possible to determine a consensus for all items of the sample and there is perfect agreement in the group of raters on each item. The proposed method is superior the consensus approach in the sense that no decision has to be made if there is no consensus in the group. Moreover, the new approach takes into account the variability in the group while the strength of consensus is not taken into account with 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 substantially according to the definition of the consensus used. The proposed kappa coefficient thus provides an alternative to the common approach which consists in summarizing the responses given by the raters in the group into a single response (the consensus) and generalizes the agreement index proposed by Schouten (1982). Further, it has the advantage of using more information than the consensus method (variability in the group of raters), of solving the problem of items without consensus and of being built upon less stringent assumptions. Experts can fix levels to interpret the values taken by the new coefficient and determine a lower bound under which the isolated rater may be rejected as in the SCT selection process or considered as "out of range".

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. 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.

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

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.

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.

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.