British Journal of Mathematical and Statistical Psychology (2008), 61, 29–48

The British Psychological Society

www.bpsjournals.co.uk

Computing inter-rater reliability and its variance in the presence of high agreement

Kilem Li Gwet*
STATAXIS Consulting, Gaithersburg, USA

Pi (π) and kappa (κ) statistics are widely used in the areas of psychiatry and psychological testing to compute the extent of agreement between raters on nominally scaled data. It is a fact that these coefficients occasionally yield unexpected results in situations known as the paradoxes of kappa. This paper explores the origin of these limitations, and introduces an alternative and more stable agreement coefficient referred to as the AC_1 coefficient. Also proposed are new variance estimators for the multiple-rater generalized π and AC_1 statistics, whose validity does not depend upon the hypothesis of independence between raters. This is an improvement over existing alternative variances, which depend on the independence assumption. A Monte-Carlo simulation study demonstrates the validity of these variance estimators for confidence interval construction, and confirms the value of AC_1 as an improved alternative to existing inter-rater reliability statistics.

1. Introduction

Researchers in various fields often need to evaluate the quality of a data collection method. In many studies, a data collection tool, such as a survey questionnaire, a laboratory procedure or a classification system, is used by different people referred to as raters, observers or judges. In an effort to minimize the effect of the rater factor on data quality, investigators like to know whether all raters apply the data collection method in a consistent manner. Inter-rater reliability quantifies the closeness of scores assigned by a pool of raters to the same study participants. The closer the scores, the higher the reliability of the data collection method. Although reliability data can be discrete or continuous, the focus of this paper is on inter-rater reliability assessment on nominally scaled data. Such data are typically obtained from studies where raters must classify study participants into one category among a limited number of possible categories.

Banerjee, Capozzoli, McSweeney, and Sinha (1999) provide a good review of the techniques developed to date for analysing nominally scaled data. Two of the most influential papers in this area are those of Fleiss (1971) and Fleiss, Cohen, and Everitt

*Correspondence should be addressed to Dr Kilem Li Gwet, Statistical Consultant, STATAXIS Consulting, PO Box 2696, Gaithersburg, MD 20886-2696, USA (e-mail: gwet62@hotmail.com).

DOI:10.1348/000711006X126600

(1969), which contain the most popular results in use today. Fleiss *et al.* provide large sample approximations of the variances of the κ and weighted κ statistics suggested by Cohen (1960, 1968), respectively in the case of two raters, while Fleiss extends the π-statistic to the case of multiple raters. Landis and Koch (1977) also give an instructive discussion of inter-rater agreement among multiple observers. Agresti (2002) presents several modelling techniques for analysing rating data in addition to presenting a short account of the state of the art. Light (1971) introduces measures of agreement conditionally upon a specific classification category, and proposes a generalization of Cohen's κ-coefficient to the case of multiple raters. Conger (1980) suggests an alternative multiple-rater agreement statistic obtained by averaging all pairwise overall and chance-corrected probabilities proposed by Cohen (1960). Conger (1980) also extends the notion of pairwise agreement to that of *g*-wise agreement where agreement occurs if *g* raters rather than two raters classify an object into the same category.

In section 2, I introduce the most commonly used pairwise indexes. Section 3 discusses a theoretical framework for analysing the origins of the kappa's paradox. An alternative and more stable agreement coefficient referred to as the AC_1 statistic is introduced in section 4. Section 5 is devoted to the analysis of the bias associated with the various pairwise agreement coefficients under investigation. In section 6, a variance estimator for the generalized π-statistic is proposed, which is valid even under the assumption of dependence of ratings. Section 7 presents a variance estimator of the AC_1 statistic, which is always valid. The important special case of two raters is discussed in section 8, while section 9 describes a small simulation study aimed at verifying the validity of the variance estimators as well as the magnitude of the biases associated with the various indexes under investigation.

2. Cohen's κ, Scott's π, *G*-index and Fleiss's generalized π

In a two-rater reliability study involving raters *A* and *B*, the data will be reported in a two-way contingency table such as Table 1. Table 1 shows the distribution of n study participants by rater and response category, where n_{kl} indicates the number of participants that raters *A* and *B* classified into categories k and l, respectively.

All inter-rater reliability coefficients discussed in this paper have two components: the overall agreement probability p_a, which is common to all coefficients, and the chance-agreement probability p_e, which is specific to each index. For the two-rater

Table 1. Distribution of n participants by rater and response category

	Rater *B*				
Rater *A*	1	2	⋯	q	Total
1	n_{11}	n_{12}	⋯	n_{1q}	n_{A1}
2	n_{21}	n_{22}	⋯	n_{2q}	n_{A2}
⋮			⋯		⋮
q	n_{q1}	n_{q2}	⋯	n_{qq}	n_{Aq}
Total	n_{B1}	n_{B2}	⋯	n_{Bq}	n

reliability data of Table 1, the overall agreement probability is given by:

$$p_{\mathrm{a}} = \sum_{k=1}^{q} p_{kk}, \quad \text{where} \quad p_{kk} = n_{kk}/n.$$

Let $p_{Ak} = n_{Ak}/n$, $p_{Bk} = n_{Bk}/n$, and $\hat{\pi}_k = (p_{Ak} + p_{Bk})/2$. Cohen's κ-statistic is given by:

$$\hat{\gamma}_{\kappa} = (p_{\mathrm{a}} - p_{\mathrm{e}|\kappa})/(1 - p_{\mathrm{e}|\kappa}), \quad \text{where} \quad p_{\mathrm{e}|\kappa} = \sum_{k=1}^{q} p_{Ak} p_{Bk}.$$

Scott (1955) proposed the π-statistic given by:

$$\hat{\gamma}_{\pi} = (p_{\mathrm{a}} - p_{\mathrm{e}|\pi})/(1 - p_{\mathrm{e}|\pi}), \quad \text{where} \quad p_{\mathrm{e}|\pi} = \sum_{k=1}^{q} \hat{\pi}_k^2.$$

The *G*-index of Holley and Guilford (1964) is given by:

$$\hat{\gamma}_G = (p_{\mathrm{a}} - p_{\mathrm{e}|G})/(1 - p_{\mathrm{e}|G}),$$

where $p_{\mathrm{e}|G} = 1/q$, and q represents the number of response categories. Note that the expression used for $\hat{\gamma}_G$ here is more general than the original Holley–Guilford formula, which was presented for the simpler situation of two raters and two response categories only.

If a reliability study involves an arbitrarily large number r of raters, rating data are often reported in a frequency table showing the distribution of raters by participant and response category, as described in Table 2. For a given participant i and category k, r_{ik} represents the number of raters who classified participant i into category k.

Table 2. Distribution of r raters by participant and response category

	Category				
Participant	1	2	⋯	q	Total
1	r_{11}	r_{12}	⋯	r_{1q}	r
2	r_{21}	r_{22}	⋯	r_{2q}	r
⋮			⋯		⋮
n	r_{n1}	r_{n2}	⋯	r_{nq}	r
Total	r_{+1}	r_{+2}	⋯	r_{+q}	nr

Fleiss (1971) extended Scott's π-statistic to the case of multiple raters (r) and proposed the following equation:

$$\hat{\gamma}_{\pi} = \frac{p_{\mathrm{a}} - p_{\mathrm{e}|\pi}}{1 - p_{\mathrm{e}|\pi}}, \quad \text{where} \quad \begin{cases} p_{\mathrm{a}} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \sum_{k=1}^{q} \frac{r_{ik}(r_{ik} - 1)}{r(r-1)}, \\ p_{\mathrm{e}|\pi} = \displaystyle\sum_{k=1}^{q} \hat{\pi}_k^2, \ \text{and} \ \hat{\pi}_k = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \frac{r_{ik}}{r}. \end{cases} \tag{1}$$

The terms p_a and $p_{e|\pi}$ are, respectively, the overall agreement probability and the probability of agreement due to chance. Conger (1980) suggested a generalized version of the κ-statistic that is obtained by averaging all $r(r-1)/2$ pairwise κ-statistics as defined by Cohen (1960). The κ-statistic can also be generalized as follows:

$$\hat{\gamma}_\kappa = \frac{p_a - p_{e|\kappa}}{1 - p_{e|\kappa}},$$

where p_a is defined as above and chance-agreement probability $p_{e|\kappa}$ given by:

$$p_{e|\kappa} = \sum_{k=1}^{q} \sum_{\alpha=2}^{r} (-1)^\alpha \left(\sum_{i_1 < \cdots < i_\alpha} \prod_{j=1}^{\alpha} p_{ki_j} \right). \tag{2}$$

The term $p_{ki_j}(j = 1, ..., \alpha)$ represents the proportion of participants that rater i_j classified into category k. It follows from equation (2) that if $r = 2$, then $p_{e|\kappa}$ reduces to the usual formula of chance-agreement probability for the κ-statistic. For $r = 3$ and $r = 4$ the chance-agreement probabilities are, respectively, given by:

$$p_{e|\kappa}(3) = \sum_{k=1}^{q} (p_{k1}p_{k2} + p_{k1}p_{k3} + p_{k2}p_{k3} - p_{k1}p_{k2}p_{k3}),$$

$$\begin{aligned} p_{e|\kappa}(4) = \sum_{k=1}^{q} [&(p_{k1}p_{k2} + p_{k1}p_{k3} + p_{k1}p_{k4} + p_{k2}p_{k3} + p_{k2}p_{k4} + p_{k3}p_{k4}) \\ &- (p_{k1}p_{k2}p_{k3} + p_{k1}p_{k2}p_{k4} + p_{k1}p_{k3}p_{k4} + p_{k2}p_{k3}p_{k4}) \\ &+ p_{k1}p_{k2}p_{k3}p_{k4}]. \end{aligned}$$

This general version of the κ-statistic has not been studied yet and no expression for its variance is available. There is no indication, however, that it has better statistical properties than Fleiss's generalized statistic. Nevertheless, a practitioner interested in using this estimator, may still estimate its variance using the jackknife method described by equation (36) for the π-statistic. Hubert (1977) discusses other possible extensions of the κ-statistic to the case of multiple raters.

3. Paradox's origin

Table 3 contains an example of rating data. These illustrate the limitations of equation (1) as a measure of the extent of agreement between raters. For those data,

Table 3. Distribution of 125 participants by rater and response category

	Rater *B*		
Rater *A*	+	−	Total
+	118	5	123
−	2	0	2
Total	120	5	125

$\hat{\gamma}_\pi = (0.9440 - 0.9456)/(1 - 0.9456) = -0.0288$, which is even a negative value. This result is the opposite of what our intuition would suggest and illustrates one of the paradoxes noted by Cicchetti and Feinstein (1990) where high agreement is coupled with low κ. In this example, raters A and B are expected to have a high inter-rater reliability.

To understand the nature and the causes of the paradoxical behaviour of the π- and κ-statistics, I will confine myself to the case of two raters, A and B, who must identify the presence or absence of a trait on individuals of a given population of interest. These individuals will eventually be selected to participate in a study, and are therefore potential study participants. The two raters will classify participants into the ' + ' or ' − ' categories according to whether the trait is found or not. I will study how agreement indexes are affected by raters' sensitivity, specificity and the trait prevalence in the population. The rater's sensitivity is defined as the conditional probability of classifying a participant into the ' + ' category given that the trait is indeed present. The rater's specificity is the conditional probability of classifying a participant into the ' − ' category given that the trait is actually absent.

Let α_A and α_B denote, respectively, raters A and B sensitivity values. Similarly, β_A and β_B will denote raters A and B specificity values. It follows that the probabilities P_{A+} and P_{B+} for raters A and B to classify a participant into the ' + ' category are given by

$$P_{A+} = P_r\alpha_A + (1 - P_r)(1 - \beta_A), \tag{3}$$

$$P_{B+} = P_r\alpha_B + (1 - P_r)(1 - \beta_B), \tag{4}$$

where P_r represents the population trait prevalence. Our objective is to study how trait prevalence, sensitivity and specificity affect inter-rater reliability. For the sake of simplicity I will make the following two assumptions:

(*A*1) Sensitivity and specificity are identical for both raters. That is $\alpha_A = \beta_A$ and $\alpha_B = \beta_B$.

(*A*2) Correct classifications are independent. That is, if α_{AB} denotes the probability that raters A and B correctly classify an individual into the ' + ' category, then $\alpha_{AB} = \alpha_A\alpha_B$.

The probability P_a that both raters agree is given by $P_a = \pi_{++} + \pi_{--}$, where π_{++} and π_{--} are obtained as follows

$$\pi_{++} = \alpha_A\alpha_B P_r + (1 - P_r)(1 - \beta_A)(1 - \beta_B) \quad = \quad \alpha_A\alpha_B P_r + (1 - P_r)(1 - \alpha_A)(1 - \alpha_B),$$

and $\pi_{--} = 1 - (P_{A+} + P_{B+} - \pi_{++})$. The following important equation can be established:

$$P_a = (1 - \alpha_A)(1 - \alpha_B) + \alpha_A\alpha_B. \tag{5}$$

This relation shows that the overall agreement probability between two raters A and B does not depend upon the trait prevalence. Rather, it depends upon the rater's sensitivity and specificity values.

The partial derivative with respect to P_r of an inter-rater coefficient of the form $\gamma = (P_a - P_e)/(1 - P_e)$ is given by

$$\partial\gamma/\partial P_r = -\frac{1 - P_a}{(1 - P_e)^2}\partial P_e/\partial P_r, \tag{6}$$

since, from equation (5), one can conclude that $\partial P_a/\partial P_r = 0$. For Scott's π and Cohen's κ-statistics,

$$\partial\gamma_\pi/\partial P_r = 2\frac{(1 - 2\lambda)^2(1 - P_a)(1 - 2P_r)}{(1 - P_{e|\pi})^2}, \tag{7}$$

$$\partial\gamma_\kappa/\partial P_r = 2\frac{(1 - 2\alpha_A)(1 - 2\alpha_B)(1 - P_a)(1 - 2P_r)}{(1 - P_{e|\kappa})^2}, \tag{8}$$

where $\lambda = (\alpha_A + \alpha_B)/2$. Let π_+ be the probability that a randomly chosen rater classifies a randomly chosen participant into the ' + ' category. Then,

$$\pi_+ = (P_{A+} + P_{B+})/2 = \lambda P_r + (1 - \lambda)(1 - P_r). \tag{9}$$

The two equations (7) and (8) are derived from the fact that $P_{e|\pi} = \pi_+^2 + (1 - \pi_+)^2$ and $P_{e|\kappa} = 1 - (P_{A+} + P_{B+}) + 2P_{A+}P_{B+}$. It follows from assumption $A1$ that $\partial\gamma_G/\partial P_r = 0$, since $\hat{\gamma}_G$ is solely a function of P_a. Under this assumption, the G-index takes a constant value of $2P_a - 1$ that depends on the raters' sensitivity. Equation (6) shows that chance-agreement probability plays a pivotal role on how inter-rater reliability relates to trait prevalence. Equation (7) indicates that Scott's π-statistic is an increasing function of P_r for the values of trait prevalence between 0 and 0.50, and becomes decreasing for $P_r > 0.50$, reaching its maximum value when $P_r = 0.50$. Because $0 \leq P_r \leq 1$, $\hat{\gamma}_\pi$ takes its smallest value at $P_r = 0$ and $P_r = 1$. Using equation (5) and the expression of $P_{e|\pi}$, one can show that:

$$\gamma_\pi = \frac{(2\lambda - 1)^2 P_r(1 - P_r) - (\alpha_A - \alpha_B)^2/4}{(2\lambda - 1)^2 P_r(1 - P_r) + \lambda(1 - \lambda)}. \tag{10}$$

It follows that,

$$\text{if } P_r = 0 \text{ or } P_r = 1 \text{ then } \gamma_\pi = -\frac{(\alpha_A - \alpha_B)^2}{4(1 - \lambda)}, \tag{11}$$

$$\text{if } P_r = 0.50 \text{ then } \hat{\gamma}_\pi = 2P_a - 1 = 1 - 4\lambda + 4\alpha_A\alpha_B, \tag{12}$$

Equations (10), (11) and (12) show very well how paradoxes often occur in practice. From equation (11) it appears that whenever a trait is very rare or omnipresent, Scott's π-statistic yields a negative inter-rater reliability regardless of the raters' sensitivity values. In other words, if prevalence is low or high, any large extent of agreement between raters will not be reflected in the π-statistic.

Equation (8), on the other hand, indicates that when trait prevalence is smaller than 0.5, Cohen's κ-statistic may be an increasing or a decreasing function of trait prevalence depending on raters A and B sensitivity values. That is, if one rater has a sensitivity smaller than 0.5 and the other a sensitivity greater than 0.5 then κ-statistic is a decreasing function of P_r, otherwise it is increasing. The situation is similar when trait prevalence is greater than 0.50. The maximum or minimum value of κ is reached at $P_r = 0.50$. If one rater has a sensitivity of 0.50, then $\kappa = 0$ regardless of the trait

prevalence. The general equation of Cohen's κ-statistic is given by:

$$\gamma_\kappa = \frac{(2\alpha_A - 1)(2\alpha_B - 1)P_r(1 - P_r)}{(2\alpha_A - 1)(2\alpha_B - 1)P_r(1 - P_r) + (1 - P_a)/2}. \quad (13)$$

It follows that:

$$\text{if } P_r = 0 \text{ or } P_r = 1 \text{ then } \hat{\gamma}_\kappa = 0, \quad (14)$$

$$\text{if } P_r = 0.50 \text{ then } \hat{\gamma}_\kappa = 2P_a - 1 = 1 - 4\lambda + 4\alpha_A\alpha_B. \quad (15)$$

Similar to Scott's π-statistic, κ seems to yield reasonable values only when trait prevalence is close to 0.5. A value of trait prevalence that is either close to 0 or close to 1 will considerably reduce the ability of κ to reflect any extent of agreement between raters.

Many inter-rater agreement coefficients proposed in the literature have been criticized on the grounds that they are dependent upon trait prevalence. Such a dependence is inevitable if raters' sensitivity levels are different from their specificity levels. In fact, without assumption *A*1, even the overall agreement probability P_a is dependent upon trait prevalence P_r due to the fact that P_a can be expressed as follows:

$$P_a = 2((\alpha_A\alpha_B - \beta_A\beta_B) - [(\alpha_A + \alpha_B) - (\beta_A + \beta_B)])P_r + (1 + 2\beta_A\beta_B - (\beta_A + \beta_B)).$$

However, the impact of prevalence on the overall agreement probability is small if sensitivity and specificity are reasonably close.

The previous analysis indicates that the *G*-index, π-statistic and κ-statistic all have the same reasonable behaviour when trait prevalence P_r takes a value in the neighbourhood of 0.5. However, their behaviour becomes very erratic (with the exception of *G*-index) as soon as trait prevalence goes to the extremes. I argue that the chance-agreement probability used in these statistics is ill-estimated when trait prevalence is in the neighbourhood of 0 or 1. I will now propose a new agreement coefficient that will share the common reasonable behaviour of its competitors in the neighbourhood of 0.5, but will outperform them when trait prevalence goes to the extremes.

4. An alternative agreement statistic

Before I introduce an improved alternative inter-rater reliability coefficient, it is necessary to develop a clear picture of the goal one normally attempts to achieve by correcting inter-rater reliability for chance agreement. My premises are the following:

(a) Chance agreement occurs when at least one rater rates an individual randomly.
(b) Only an unknown portion of the observed ratings is subject to randomness.

I will consider that a rater *A* classifies an individual into one of two categories either randomly, when he or she does not know where it belongs, or with certainty, when he or she is certain about its 'true' membership. Rater *A* performs a random rating not all the time, but with a probability θ_A. That is, θ_A is the propensity for rater *A* to perform a random rating. The participants not classified randomly are supposed to have been classified into the correct category. If the random portion of the study was identifiable, rating data of two raters *A* and *B* classifying *N* individuals into categories '+' and '−' could be reported as shown in Table 4.

Table 4. Distribution of *N* participants by rater, randomness of classification and response category

			Rater *B*			
			Random (R)		Certain (C)	
			+	−	+	−
Rater *A*	Random (R)	+	$N_{++\cdot RR}$	$N_{+-\cdot RR}$	$N_{++\cdot RC}$	$N_{+-\cdot RC}$
		−	$N_{-+\cdot RR}$	$N_{--\cdot RR}$	$N_{-+\cdot RC}$	$N_{--\cdot RC}$
	Certain (C)	+	$N_{++\cdot CR}$	$N_{+-\cdot CR}$	$N_{++\cdot CC}$	**0**
		−	$N_{-+\cdot CR}$	$N_{--\cdot CR}$	**0**	$N_{--\cdot CC}$

Note that $N_{+-\cdot RC}$ for example, represents the number of individuals that rater A classified randomly into the ' + ' category and that rater B classified with certainty into the ' − ' category. In general, for $(k, l) \in \{+, -\}$, and $(X, Y) \in \{R, C\}$, $N_{kl\cdot XY}$ represents the number of individuals that rater A classified into category k using a classification method X (random or certainty), and that rater B classified into category l using a classification method Y (random or certainty).

To evaluate the extent of agreement between raters A and B from Table 4, what is needed is the ability to remove from consideration all agreements that occurred by chance; that is $N_{++\cdot RR} + N_{++\cdot CR} + N_{++\cdot RC} + N_{--\cdot RR} + N_{--\cdot CR} + N_{--\cdot RC}$. This yields the following 'true' inter-rater reliability:

$$\gamma = \frac{N_{++\cdot CC} + N_{--\cdot CC}}{\sum_{k\in\{+,-\}} \left(N_{kk\cdot RR} + N_{kk\cdot CR} + N_{kk\cdot RC}\right)} \tag{16}$$

Equation (16) can also be written as:

$$\gamma = \frac{P_a - P_e}{1 - P_e}, \quad \text{where}$$
$$P_a = \sum_{k\in\{+,-\}} \sum_{X,Y\in\{C,R\}} \frac{N_{kk\cdot XY}}{N}, \quad \text{and} \quad P_e = \sum_{k\in\{+,-\}} \sum_{\substack{X,Y\in\{C,R\} \\ (X,Y)\neq(C,C)}} \frac{N_{kk\cdot XY}}{N} \tag{17}$$

In a typical reliability study the two raters A and B would rate n study participants, and rating data reported as shown in Table 1, with $q = 2$. The problem is to find a good statistic $\hat{\gamma}$ for estimating γ. A widely accepted statistic for estimating the overall agreement probability P_a is given by:

$$p_a = (n_{++} + n_{--})/n. \tag{18}$$

The estimation of P_e represents a more difficult problem, since it requires one to be able to isolate ratings performed with certainty from random ratings. To get around this difficulty, I decided to approximate P_e by a parameter that can be quantified more easily, and to evaluate the quality of the approximation in section 5.

Suppose an individual is selected randomly from a pool of individuals and rated by raters A and B. Let G and R be two events defined as follows:

$$G = \{\text{The two raters } A \text{ and } B \text{ agree}\}, \tag{19}$$

$$R = \{\text{A rater } (A, \text{ or } B, \text{ or both}) \text{ performs a random rating}\}. \tag{20}$$

It follows that $P_e = P(G \cap R) = P(G/R)P(R)$, where $P(G/R)$ is the conditional probability that A and B agree given that one of them (or both) has performed a random rating.

A random rating would normally lead to the classification of an individual into either category with the same probability 1/2, although this may not always be case. Since agreement may occur on either category, it follows that $P(G/R) = 2 \times 1/2^2 = 1/2$. As for the estimation of the probability of random rating $P(R)$, one should note that when the trait prevalence P_r is high or low (i.e. if $P_r(1 - P_r)$ is small), a uniform distribution of participants among categories is an indication of high proportion of random ratings, hence of high probability $P(R)$.

Let the random variable X_+ be defined as follows:

$$X_+ = \begin{cases} 1 & \text{if a rater classifies the participant into category } +, \\ 0 & \text{otherwise.} \end{cases}$$

I suggest approximating $P(R)$ with a normalized measure of randomness Ψ defined by the ratio of the variance $V(X_+)$ of X_+ to the maximum possible variance V_{MAX} for X_+, which is reached only when the rating is totally random. It follows that

$$\Psi = V(X_+)/V_{MAX} = \frac{\pi_+(1 - \pi_+)}{1/2(1 - 1/2)} = 4\pi_+(1 - \pi_+),$$

where π_+ represents the probability that a randomly chosen rater classifies a randomly chosen individual into the ' + ' category. This leads to the following formulation of chance agreement:

$$P_e^* = P(G/R)\Psi = 2\pi_+(1 - \pi_+). \tag{21}$$

This approximation leads to the following approximated 'true' inter-rater reliability:

$$\gamma^* = \frac{P_a - P_e^*}{1 - P_e^*}, \tag{22}$$

The probability π_+ can be estimated from sample data by $\hat{\pi}_+ = (p_{A+} + p_{B+})/2$, where $p_{A+} = n_A + /n$ and $p_{B+} = n_{B+}/n$. This leads to a chance-agreement probability estimator $p_e^* = P(G/R)\hat{\Psi}$, where $\hat{\Psi} = 4\hat{\pi}_+(1 - \hat{\pi}_+)$. That is,

$$p_e^* = 2\hat{\pi}_+(1 - \hat{\pi}_+). \tag{23}$$

Note that $\hat{\pi}_+(1 - \hat{\pi}_+) = \hat{\pi}_-(1 - \hat{\pi}_-)$. Therefore, p_e^* can be rewritten as $p_e^* = \hat{\pi}_+(1 - \hat{\pi}_+) + \hat{\pi}_-(1 - \hat{\pi}_-)$.

The resulting agreement statistic is given by,

$$\hat{\gamma}_1 = (p_a - p_e^*)/(1 - p_e^*),$$

with p_a given by equation (18), and is shown in section 5 mathematically to have a smaller bias with respect to the 'true' agreement coefficient than all its competitors.

Unlike the κ- and π-statistics, this agreement coefficient uses a chance-agreement probability that is calibrated to be consistent with the propensity of random rating that is suggested by the observed ratings. I will refer to the calibrated statistic $\hat{\gamma}_1$ as the AC_1 estimator, where AC stands for agreement coefficient and digit 1 indicates the first-order chance correction, which accounts for full agreement only as opposed to full and partial agreement (second-order chance correction); the latter problem, which will be investigated elsewhere, will lead to the AC_2 statistic.

A legitimate question to be asked is whether the inter-rater reliability statistic $\hat{\gamma}_1$, estimates the 'true' inter-rater reliability of equation (16) at all, and under what circumstances. I will show in the next section that if trait prevalence is high or low, then $\hat{\gamma}_1$ does estimate the 'true' inter-rater reliability very well. However, with trait prevalence at the extremes, π, κ and *G*-index are all biased for estimating the 'true' inter-rater reliability under any circumstances.

5. Biases of inter-rater reliability statistics

Let us consider two raters, *A* and *B*, who would perform a random rating with probabilities θ_A and θ_B, respectively. Each classification of a study participant by a random mechanism will either lead to a disagreement or to an agreement by chance.

The rater's sensitivity values (which are assumed to be identical to their specificity values) are given by:

$$\alpha_A = 1 - \theta_A/2 \text{ and } \alpha_B = 1 - \theta_B/2.$$

These equations are obtained under the assumption that any rating that is not random will automatically lead to a correct classification, while a random rating leads to a correct classification with probability 1/2. In fact, $\alpha_A = (1 - \theta_A) + \theta_A/2 = 1 - \theta_A/2$.

Under this simple rating model, and following equation (5), the overall agreement probability is given by $P_a = \alpha_A\alpha_B + (1 - \alpha_A)(1 - \alpha_B) = 1 - (\theta_A + \theta_B)/2 + \theta_A\theta_B/2$. As for chance-agreement probability P_e let R_A and R_B be two events defined as follows:

- R_A: Rater *A* performs a random rating.
- R_B: Rater *B* performs a random rating.

Then,

$$P_e = P(G \cap R) = P(G \cap R_A \cap \bar{R}_B) + P(G \cap R_A \cap R_B) + P(G \cap \bar{R}_A \cap R_B)$$

$$= \theta_A(1 - \theta_B)/2 + \theta_A\theta_B/2 + \theta_B(1 - \theta_A)/2 \;=\; (\theta_A + \theta_B - \theta_A\theta_B)/2.$$

The 'true' inter-rater reliability is then given by:

$$\gamma = 2\frac{(1 - \theta_A)(1 - \theta_B)}{1 + (1 - \theta_A)(1 - \theta_B)}. \tag{24}$$

The theoretical agreement coefficients will now be derived for the AC_1, *G*-index, κ, and π statistics. Let $\lambda = 1 - (\theta_A + \theta_B)/4$.

For AC_1 coefficient, it follows from equations (5) and (21) that chance-agreement probability P_e^* is obtained as follows:

$$P_e^* = 2\pi_+(1-\pi_+) \quad = 2[\lambda P_r + (1-\lambda)(1-P_r)][(1-\lambda P_r) - (1-\lambda)(1-P_r)]$$

$$= 2\lambda(1-\lambda) + 2(1-2\lambda)^2 P_r(1-P_r) \quad = P_e - (\theta_A - \theta_B)^2/8 + \Delta,$$

where $\Delta = 2(1-2\lambda)^2 P_r(1-P_r)$. The theoretical AC_1 coefficient is given by:

$$\gamma_1 = \gamma - (1-\gamma)\frac{(\theta_A - \theta_B)^2 - \Delta}{(1-P_e) + [(\theta_A - \theta_B)^2/8 - \Delta]}. \tag{25}$$

For Scott's π-coefficient, one can establish that the chance-agreement probability $P_{e|\pi}$ is given by $P_{e|\pi} = P_e + (1-\theta_A)(1-\theta_B) + (\theta_A - \theta_B)^2/8 - \Delta$. This leads to Scott's π-coefficient of

$$\gamma_\pi = \gamma - (1-\gamma)\frac{(1-\theta_A)(1-\theta_B) + (\theta_A - \theta_B)^2/8 - \Delta}{(1-P_e) - [(1-\theta_A)(1-\theta_B) + (\theta_A - \theta_B)^2/8] + \Delta}. \tag{26}$$

For the *G*-index, $P_{e|G} = 1/2 = P_e + (1-\theta_A)(1-\theta_B)/2$:

$$\gamma_G = \gamma - (1-\gamma)\frac{(1-\theta_A)(1-\theta_B)/2}{(1-P_e) - (1-\theta_A)(1-\theta_B)/2}. \tag{27}$$

For Cohen's κ-coefficient, $P_{e|\kappa} = P_e + (1-\theta_A)(1-\theta_B) - \Delta_\kappa$, where $\Delta_\kappa = 2(1-\theta_A)(1-\theta_B)P_r(1-P_r)$:

$$\gamma_\kappa = \gamma - (1-\gamma)\frac{(1-\theta_A)(1-\theta_B) - \Delta_\kappa}{(1-P_e) - [(1-\theta_A)(1-\theta_B) - \Delta_\kappa]}. \tag{28}$$

To gain further insight into the magnitude of the biases of these different inter-rater reliability statistics, let us consider the simpler case where raters *A* and *B* have the same propensity for random rating; that is, $\theta_A = \theta_B = \theta$. The 'true' inter-rater reliability is given by:

$$\gamma = \frac{2(1-\theta)^2}{1+(1-\theta)^2}. \tag{29}$$

I define the bias of an agreement coefficient γ_X as $B_X(\theta) = \gamma_X - \gamma$, the difference between the agreement coefficient and the 'true' coefficient. The biases of AC_1, π, κ and *G*-index statistics, respectively denoted by $B_1(\theta)$, $B_\pi(\theta)$, $B_\kappa(\theta)$ and $B_G(\theta)$, satisfy the following relations:

$$B_G(\theta) = -\frac{\theta(1-\theta)^2(2-\theta)}{1+(1-\theta)^2}, \quad -\frac{\theta(1-\theta)^2(2-\theta)}{1+(1-\theta)^2} \le B_1(\theta) \le 0,$$

$$-2\frac{(1-\theta)^2}{1+(1-\theta)^2} \le B_\pi(\theta) \le -\frac{\theta(1-\theta)^2(2-\theta)}{1+(1-\theta)^2},$$

$$-2\frac{(1-\theta)^2}{1+(1-\theta)^2} \le B_\kappa(\theta) \le -\frac{\theta(1-\theta)^2(2-\theta)}{1+(1-\theta)^2}.$$

Which way the bias will go depends on the magnitude of trait prevalence. It follows from these equations that the *G*-index consistently exhibits a negative bias, which may

take a maximum absolute value around 17%, when the rater's propensity for random rating is around 35%, and will gradually decrease as θ goes to 1. The AC_1 statistic, on the other hand, has a negative bias that ranges from $-\theta(1-\theta)^2(2-\theta)/(1+(1-\theta)^2)$ to 0, reaching its largest absolute value of $\theta(1-\theta)^2(2-\theta)/(1+(1-\theta)^2)$ only when the trait prevalence is around 50%. The remaining two statistics, on the other hand, have some serious bias problems on the negative side. The π and κ statistics each have a bias whose lowest value is $-2(1-\theta)^2/[1+(1-\theta)^2]$, which varies from 0 to −1. That means π and κ may underestimate the 'true' inter-rater reliability by 100%.

The next two sections, 6 and 7, are devoted to variance estimation of the generalized π-statistic and the AC_1 statistic, respectively, in the context of multiple raters. For the two sections, I will assume that the n participants in the reliability study were randomly selected from a bigger population of N potential participants. Likewise, the r raters can be assumed to belong to a bigger universe of R potential raters. This finite-population framework has not yet been considered in the study of inter-rater agreement assessment. For this paper, however, I will confine myself to the case where $r = R$, that is the estimators are not subject to any variability due to the sampling of raters. Methods needed to extrapolate to a bigger universe of raters will be discussed in a different paper.

6. Variance of the generalized π-statistic

The π-statistic denoted by $\hat{\gamma}_\pi$ is defined as follows:

$$\hat{\gamma}_\pi = \frac{p_a - p_{e|\pi}}{1 - p_{e|\pi}}, \tag{30}$$

where p_a and $p_{e|\pi}$ are defined as follows:

$$p_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r(r-1)}, \text{ and } p_{e|\pi} = \sum_{k=1}^{q}\hat{\pi}_k^2, \text{ with } \hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r}. \tag{31}$$

Concerning the estimation of the variance of $\hat{\gamma}_\pi$, Fleiss (1971) suggested the following variance estimator under the hypothesis of no agreement between raters beyond chance:

$$v(\hat{\gamma}_\pi|\text{No agreement}) = \frac{2(1-f)}{nr(r-1)} \times \frac{p_{e|\pi} - (2r-3)p_{e|\pi}^2 + 2(r-2)\sum_{k=1}^{q}\hat{\pi}_k^3}{(1-p_{e|\pi})^2}, \tag{32}$$

where $f = n/N$ is the sampling fraction, which could be neglected if the population of potential participants is deemed very large. It should be noted that this variance estimator is invalid for confidence interval construction. The original expression proposed by Fleiss does not include the finite-population correction factor $1-f$. Cochran (1977) is a good reference for readers interested in statistical methods in finite-population sampling.

I propose here a non-paramateric variance estimator for $\hat{\gamma}_\pi$ that is valid for confidence interval construction using the linearization technique. Unlike

$v(\hat{\gamma}_\pi|\text{No agreement})$, the validity of the non-parametric variance estimator does not depend on the extent of agreement between the raters. This variance estimator is given by

$$v(\hat{\gamma}_\pi) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\gamma}^*_{\pi i} - \hat{\gamma}_\pi)^2, \tag{33}$$

where $\hat{\gamma}^*_{\pi i}$ is given by

$$\hat{\gamma}^*_{\pi i} = \hat{\gamma}_{\pi i} - 2(1-\hat{\gamma}_\pi)\frac{p_{e\pi|i} - p_{e|\pi}}{1-p_{e|\pi}}, \tag{34}$$

where $\hat{\gamma}_{\pi i} = (p_{a|i} - p_{e|\pi})/(1-p_{e|\pi})$, and $p_{a|i}$, and $p_{e\pi|i}$ are given by:

$$p_{a|i} = \sum_{i=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r(r-1)}, \text{ and } p_{e\pi|i} = \sum_{k=1}^{q}\frac{r_{ik}}{r}\hat{\pi}_k. \tag{35}$$

To see how equation (33) is derived, one should consider the standard approach that consists of deriving an approximation of the actual variance of the estimator and using a consistent estimator of that approximate variance as the variance estimator. Let us assume that as the sample size n increases, the estimated chance-agreement probability $p_{e|\pi}$ converges to a value $P_{e|\pi}$ and that each $\hat{\pi}_k$ converges to π_k. If $\hat{\pi}$ and π denote the vectors of the $\hat{\pi}_k$'s and π_k's, respectively, it can be shown that,

$$p_{e|\pi} - P_{e|\pi} = \frac{2}{n}\sum_{i=1}^{n}(p_{e\pi|i} - P_{e|\pi}) + O_p(\|\hat{\pi} - \pi\|^2),$$

and that if $G_\pi = (p_a - P_{e|\pi})/(1-P_{e|\pi})$, then $\hat{\gamma}_\pi$ can be expressed as,

$$\hat{\gamma}_\pi = \frac{(p_a - P_{e|\pi}) - (1-G_\pi)(p_{e|\pi} - P_{e|\pi})}{1-P_{e|\pi}} + O_p((\,p_{e|\pi} - P_{e|\pi})^2).$$

The combination of these two equations gives us an approximation of $\hat{\gamma}_\pi$ that is a linear function of r_{ik} and that captures all terms except those with a stochastic order of magnitude of $1/n$, which can be neglected. Bishop, Fienberg, and Holland (1975, chapter 14) provide a detailed discussion of the concept of stochastic order of magnitude.

The variance estimator of equation (33) can be used for confidence interval construction as well as for hypothesis testing. Its validity is confirmed by the simulation study presented in section 9.

Alternatively, a jackknife variance estimator can be used to estimate the variance of the π-statistic. The jackknife technique introduced by Quenouille (1949) and developed by Tukey (1958), is a general purpose technique for estimating variances. It has wide applicability although it is computation intensive. The jackknife variance of $\hat{\gamma}_\pi$ is given by:

$$v_J(\hat{\gamma}_\pi) = \frac{(1-f)(n-1)}{n}\sum_{i=1}^{n}(\hat{\gamma}^{(i)}_\pi - \hat{\gamma}^{(\bullet)}_\pi)^2, \tag{36}$$

where $\hat{\gamma}^{(i)}_\pi$ is the π-statistic obtained after removing participant i from the sample, while $\hat{\gamma}^{(\bullet)}_\pi$ represents the average of all $\hat{\gamma}^{(i)}_\pi$'s. Simulation results not reported in this paper show

that this jackknife variance works well for estimating the variance of $\hat{\gamma}_\pi$. The idea of using the jackknife methodology for estimating the variance of an agreement coefficient was previously evoked by Kraemer (1980).

7. Variance of the generalized AC_1 estimator

The AC_1 statistic $\hat{\gamma}_1$ introduced in section 4 can be extended to the case of r raters ($r > 2$) and q response categories ($q > 2$) as follows:

$$\hat{\gamma}_1 = \frac{p_a - p_{e|\gamma}}{1 - p_{e|\gamma}}, \tag{37}$$

where p_a is defined in equation (1), and chance-agreement probability $p_{e|\gamma}$ defined as follows:

$$p_{e|\gamma} = \frac{1}{q-1}\sum_{k=1}^{q}\hat{\pi}_k(1-\hat{\pi}_k), \tag{38}$$

the $\hat{\pi}_k$'s being defined in equation (1).

The estimator $\hat{\gamma}_1$ is a non-linear statistic of the r_{ik}'s. To derive its variance, I have used a linear approximation that includes all terms with a stochastic order of magnitude up to $n^{-1/2}$. This will yield a correct asymptotic variance that includes all terms with an order of magnitude up to $1/n$. Although a rigorous treatment of the asymptotics is not presented here, it is possible to establish that for large values of n, a consistent estimator for estimating the variance of $\hat{\gamma}_1$ is given by:

$$v(\hat{\gamma}_1) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\gamma}^*_{1|i} - \hat{\gamma}_1)^2, \tag{39}$$

where $f = n/N$ is the sampling fraction,

$$\hat{\gamma}^*_{1|i} = \hat{\gamma}_{1|i} - 2(1-\hat{\gamma}_1)\frac{p_{e\gamma|i} - p_{e|\gamma}}{1 - p_{e|\gamma}},$$

$\hat{\gamma}_{1|i} = (p_{a|i} - p_{e|\gamma})/(1 - p_{e|\gamma})$ is the agreement coefficient with respect to participant i, $p_{a|i}$ is given by,

$$p_{a|i} = \sum_{k=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r(r-1)},$$

and chance-agreement probability with respect to unit i, $p_{e\gamma|i}$ is given by:

$$p_{e\gamma|i} = \frac{1}{q-1}\sum_{k=1}^{q}\frac{r_{ik}}{r}(1-\hat{\pi}_k),$$

To obtain equation (39), one should first derive a large-sample approximation of the actual variance of $\hat{\gamma}_1$. This is achieved by considering that as the size n of the participant sample increases, chance-agreement probability $p_{e|\gamma}$ converges to a fixed probability $P_{e|\gamma}$ and each classification probability $\hat{\pi}_k$ converges to a constant π_k. Let

us define the following two vectors: $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, ..., \hat{\pi}_q)'$ and $\boldsymbol{\pi} = (\pi_1, ..., \pi_q)'$. One can establish that:

$$p_{e|\gamma} - P_{e|\gamma} = \frac{2}{n}\sum_{i=1}^{n}(p_{e\gamma|i} - P_{e|\gamma}) + O_p(\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|),$$

$$\hat{\gamma}_1 = \frac{(p_a - P_{e|\gamma}) - (1 - G_\gamma)(p_{e|\gamma} - P_{e|\gamma})}{1 - P_{e|\gamma}} + O_p((p_{e|\gamma} - P_{e|\gamma})^2),$$

where $G_\gamma = (p_a - P_{e|\gamma})/(1 - P_{e|\gamma})$. Combining these two expressions leads to a linear approximation of $\hat{\gamma}_1$, which can be used to approximate the asymptotic variance of $\hat{\gamma}_1$.

An alternative approach for estimating the variance of $\hat{\gamma}_1$ is the jackknife method. The jackknife variance estimator is given by:

$$v_J(\hat{\gamma}_1) = (1 - f)\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\gamma}_1^{(i)} - \hat{\gamma}_1^{(\bullet)})^2, \tag{40}$$

where $\hat{\gamma}_1^{(i)}$ represents the estimator $\hat{\gamma}_1$ computed after removing participant i from the participant sample, and $\hat{\gamma}_1^{(\bullet)}$ the average of all the $\hat{\gamma}_1^{(i)}$'s.

8. Special case of two raters

Two-rater reliability studies are of special interest. Rating data in this case are often conveniently reported using the distribution of participants by rater and response category as shown in Table 1. Therefore, the inter-rater reliability coefficient and its associated variance must be expressed as functions of the n_{kl}'s.

For two raters classifying n participants into q response categories, Fleiss *et al.* (1969) proposed an estimator $v(\hat{\gamma}_\kappa|\text{No agreement})$ for estimating the variance of Cohen's κ-statistic under the hypothesis of no agreement between the raters. If there exists an agreement between the two raters, Fleiss *et al.* recommended another variance estimator $v(\hat{\gamma}_\kappa|\text{Agreement})$. These estimators are given by:

$$v(\hat{\gamma}_\kappa|\text{No agreement}) = \frac{1-f}{n(1-p_{e|\kappa})^2}\left\{\sum_{k=1}^{q} p_{Bk}\ p_{Ak}[1 - (p_{Bk} + p_{Ak})]^2 + \sum_{k=1}^{q}\sum_{\substack{l=1\\k\neq l}}^{q} p_{Bk}\ p_{Al}(p_{Bk} + p_{Al})^2 - p_{e|\kappa}^2\right\} \tag{41}$$

and

$$v(\hat{\gamma}_\kappa|\text{Agreement}) = \frac{1-f}{n(1-p_{e|\kappa})^2}\left\{\sum_{k=1}^{q} p_{kk}[1 - (p_{Ak} + p_{Bk})(1 - \hat{\gamma}_\kappa)]^2 + (1 - \hat{\gamma}_\kappa)^2\sum_{k=1}^{q}\sum_{\substack{l=1\\k\neq l}}^{q} p_{kl}(p_{Bk} + p_{Al})^2 - [\hat{\gamma}_\kappa - p_{e|\kappa}(1 - \hat{\gamma}_\kappa)]^2\right\}. \tag{42}$$

It can be shown that $v(\hat{\gamma}_\kappa|\text{Agreement})$ captures all terms of magnitude order up to n^{-1}, is consistent for estimating the true population variance and provides valid normality-based confidence intervals when the number of participants is reasonably large.

When $r = 2$, the variance of the AC_1 statistic given in equation (39) reduces to the following estimator:

$$v(\hat{\gamma}_1) = \frac{1-f}{n(1-p_{e|\gamma})^2}\left\{p_a(1-p_a) - 4(1-\hat{\gamma}_1)\left(\frac{1}{q-1}\sum_{k=1}^{q}p_{kk}(1-\hat{\pi}_k) - p_a p_{e|\gamma}\right)\right.$$
$$\left. + 4(1-\hat{\gamma}_1)^2\left(\frac{1}{(q-1)^2}\sum_{k=1}^{q}\sum_{l=1}^{q}p_{kl}[1-(\hat{\pi}_k+\hat{\pi}_l)/2]^2 - p_{e|\gamma}^2\right)\right\}. \tag{43}$$

As for Scott's π-estimator, its correct variance is given by:

$$v(\hat{\gamma}_\pi) = \frac{1-f}{n(1-p_{e|\pi})^2}\left\{p_a(1-p_a) - 4(1-\hat{\gamma}_\pi)\left(\sum_{k=1}^{q}p_{kk}\hat{\pi}_k - p_a p_{e|\pi}\right)\right.$$
$$\left. + 4(1-\hat{\gamma}_\pi)^2\left(\sum_{k=1}^{q}\sum_{l=1}^{q}p_{kl}[(\hat{\pi}_k+\hat{\pi}_l)/2]^2 - p_{e|\pi}^2\right)\right\} \tag{44}$$

For the sake of comparability, one should note that the correct variance of kappa can be rewritten as follows:

$$v(\hat{\gamma}_\kappa) = \frac{1-f}{n(1-p_{e|\kappa})^2}\left\{p_a(1-p_a) - 4(1-\hat{\gamma}_\kappa)\left(\sum_{k=1}^{q}p_{kk}\hat{\pi}_k - p_a p_{e\kappa}\right)\right.$$
$$\left. + 4(1-\hat{\gamma}_\kappa)^2\left(\sum_{k=1}^{q}\sum_{l=1}^{q}p_{kl}[(p_{Ak}+p_{Bl})/2]^2 - p_{e|\kappa}^2\right)\right\}. \tag{45}$$

The variance of the *G*-index is given by:

$$v(\hat{\gamma}_G) = 4\frac{1-f}{n}p_a(1-p_a). \tag{46}$$

Using the rating data of Table 3, I obtained the following inter-rater reliability estimates and variance estimates:

Statistic	Estimate (%)	Standard error (%)
AC_1	94.08	2.30
Kappa	–2.34	1.23
Pi	–2.88	1.09
G-Index	88.80	4.11

Because the percentage agreement p_a equals 94.4%, it appears that AC_1 and *G*-index are more consistent with the observed extent of agreement. The κ and π statistics have

low values that are very inconsistent with the data configuration and would be difficult to justify. If the standard error is compared with the inter-rater reliability estimate, the AC_1 appears to be the most accurate of all agreement coefficients.

9. Monte-Carlo simulation

In order to compare the biases of the various inter-rater reliability coefficients under investigation and to verify the validity of the different variance estimators discussed in the previous sections, I have conducted a small Monte-Carlo experiment. This experiment involves two raters, A and B, who must classify n (for $n = 20, 60, 80, 100$) participants into one of two possible categories ' + ' and ' − '.

All the Monte-Carlo experiments are based upon the assumption of a prevalence rate of $P_r = 95\%$. A propensity of random rating θ_A is set for rater A and another one θ_B for rater B at the beginning of each experiment. These parameters allow us to use equation (19) to determine the 'true' inter-rater reliability to be estimated. Each Monte-Carlo experiment is conducted as follows:

- The n participants are first randomly classified into the two categories ' + ' and ' − ' in such a way that a participant falls into category ' + ' with probability P_r.
- If a rater performs a random rating (with probabilities θ_A for rater A and θ_B for rater B), then the participant to be rated is randomly classified into one of the two categories with the same probability 1/2. A non-random rating is supposed to lead to a correct classification.
- The number of replicate samples drawn in this simulation is 500.

Each Monte-Carlo experiment has two specific objectives, which are to evaluate the magnitude of the biases associated with the agreement coefficients and to verify the validity of their variance estimators.

The bias of an estimator is measured by the difference of its Monte-Carlo expectation to the 'true' inter-rater reliability. The bias of a variance estimator, on the other hand, is obtained by comparing its Monte-Carlo expectation with the Monte-Carlo variance of the agreement coefficient. A small bias is desirable as it indicates that a given estimator or variance estimator has neither a tendency to overestimate the true population parameter nor a tendency to underestimate it.

In the simulation programmes, the calculation of the π-statistic and that of the κ-statistic were modified slightly in order to avoid the difficulty posed by undefined estimates. When $p_{e|\pi} = 1$ or $p_{e|\kappa} = 1$, these chance-agreement probabilities were replaced with 0.99999 so that the agreement coefficient can be defined.

Table 5 contains the relative bias of the agreement coefficients $\hat{\gamma}_\pi, \hat{\gamma}_\kappa, \hat{\gamma}_G$, and $\hat{\gamma}_1$. A total of 500 replicate samples were selected and for each sample s an estimate $\hat{\gamma}_s$ was calculated. The relative bias is obtained as follows:

$$\text{RelBias}(\hat{\gamma}) = \left(\frac{1}{500} \sum_{s=1}^{500} \hat{\gamma}_s - \gamma \right) / \gamma,$$

where γ is the 'true' inter-rater reliability obtained with equation (19). It follows from Table 5 that the relative bias of the AC_1 estimator, which varies from − 0.8 to 0.0% when $\theta_A = \theta_B = 5\%$, and from − 2.1 to − 1.3% when $\theta_A = 20\%$ and $\theta_B = 5\%$, is consistently smaller than the relative bias of the other inter-rater reliability statistics. The π and κ

statistics generally exhibit a very large negative bias under current conditions, ranging from −32.8 to −62.5%. The main advantage of the AC_1 statistic over the *G*-index stems from the fact that when the rater's propensity for random rating is large (i.e. around 35%), the bias of the *G*-index is at its highest, while that of the AC_1 will decrease as the trait prevalence increases.

Table 5. Relative bias of agreement coefficients for $P_r = 0.95$ based on 500 replicate samples

θ_A, θ_B	*n*	$B(\hat{\gamma}_\pi)$ %	$B(\hat{\gamma}_\kappa)$ %	$B(\hat{\gamma}_G)$ %	$B(\hat{\gamma}_1)$ %
$\theta_A^{(*)} = \theta_B^{(*)} = 5\%$	20	−32.8	−32.0	−3.6	0.0
	60	−39.5	−39.3	−5.1	−0.7
	80	−36.5	−36.4	−4.9	−0.6
	100	−35.1	−35.0	−5.2	−0.8
$\theta_A = 20\%$ $\theta_B = 5\%$	20	−62.5	−59.9	−11.9	−2.1
	60	−58.4	−57.0	−11.7	−1.4
	80	−58.2	−56.9	−12.1	−1.6
	100	−57.4	−56.3	−11.6	−1.3

(*) θ_A and θ_B represent the propensity for random rating of raters *A* and *B*, respectively.

Table 6 shows the Monte-Carlo variances of the four agreement statistics under investigation, as well as the Monte-Carlo expectations of the associated variance estimators. The Monte-Carlo expectation of a variance estimator v is obtained by averaging all 500 variance estimates v_s obtained from each replicate sample s. The Monte-Carlo variance of an agreement coefficient $\hat{\gamma}$, on the other hand, is obtained by averaging all 500 squared differences between the estimates $\hat{\gamma}_s$ and their average. More formally, the Monte-Carlo expectation $E(v)$ of a variance estimator v is defined as follows:

$$E(v) = \frac{1}{500}\sum_{s=1}^{500} v_s,$$

while the Monte-Carlo variance $V(\hat{\gamma})$ of an agreement statistic $\hat{\gamma}$ is given by:

$$V(\hat{\gamma}) = \frac{1}{500}\sum_{s=1}^{500} [\hat{\gamma}_s - \text{average}(\hat{\gamma})]^2.$$

Table 6. Monte-Carlo variances and Monte-Carlo expectations of variance estimates for $P_r = 0.95$ $\theta_A^{(*)} = \theta_B^{(*)} = 0.05$

n	$V(\hat{\gamma}_\pi)$ %	$E[v(\hat{\gamma}_\pi)]$ %	$V(\hat{\gamma}_\kappa)$ %	$E[v(\hat{\gamma}_\kappa)]$ %	$V(\hat{\gamma}_G)$ %	$E[v(\hat{\gamma}_G)]$ %	$V(\hat{\gamma}_1)$ %	$E[v(\hat{\gamma}_1)]$ %
20	15.8	3.3	15.0	3.13	0.79	0.78	0.32	0.33
60	6.0	3.9	5.9	3.83	0.28	0.31	0.10	0.12
80	3.9	3.0	3.8	3.00	0.24	0.23	0.09	0.09
100	2.5	2.4	2.5	2.39	0.17	0.19	0.07	0.07

(*) θ_A and θ_B represent the propensity for random rating of raters *A* and *B*, respectively.

It follows from Table 6 that the variance of the AC_1 statistic is smaller than that of the other statistics. In fact, $V(\hat{\gamma}_1)$ varies from 0.07% when the sample size is 100 to 0.33%

when the sample size is 20. The second smallest variance is that of the *G*-index, which varies from 0.17 to 0.79%. The κ and π statistics generally have larger variances, which range from 2% to about 15%. An examination of the Monte-Carlo expectation of the various variance estimators indicates that the proposed variance estimators for AC_1 and *G*-index work very well. Even for a small sample size, these expectations are very close to the Monte-Carlo approximations. The variance estimators of the κ and π statistics also work well except for small sample sizes, for which they underestimate the 'true' variance.

10. Concluding remarks

In this paper, I have explored the problem of inter-rater reliability estimation when the extent of agreement between raters is high. The paradox of the κ and π statistics has been investigated and an alternative agreement coefficient proposed. I have proposed new variance estimators for the κ, π and the AC_1 statistics using the linearization and jackknife methods. The validity of these variance estimators does not depend upon the assumption of independence. The absence of such variance estimators has prevented practitioners from constructing confidence intervals of multiple-rater agreement coefficients.

I have introduced the AC_1 statistic which is shown to have better statistical properties than its κ, π and *G*-index competitors. The κ and π estimators became well-known for their supposed ability to correct the percentage agreement for chance agreement. However, this paper argues that not all observed ratings would lead to agreement by chance. This will particularly be the case if the extent of agreement is high in a situation of high trait prevalence. Kappa and pi evaluate the chance-agreement probability as if all observed ratings may yield an agreement by chance. This may lead to unpredictable results with rating data that suggest a rather small propensity for chance agreement. The AC_1 statistic was developed in such a way that the propensity for chance agreement is proportional to the portion of ratings that may lead to an agreement by chance, reducing the overall agreement by chance to the right magnitude.

The simulation results tend to indicate that the AC_1 and *G*-index statistics have reasonably small biases for estimating the 'true' inter-rater reliability, while the κ and π statistics tend to underestimate it. The AC_1 outperforms the *G*-index when the trait prevalence is high or low. If the trait prevalence is around 50%, all agreement statistics perform alike. The absolute bias in this case increases with the raters' propensity for random rating, which can be reduced by giving extra training to the raters. The proposed variance estimators work well according to our simulations. For small sample sizes, the variance estimators proposed for κ and π statistics tend to underestimate the true variances.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

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

Bishop, Y. V. V., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*, 551–558.

Cochran, W. C. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

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.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*, 322–328.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*, 323–327.

Holley, J. W., & Guilford, J. P. (1964). A note on the *G* index of agreement. *Educational and Psychological Measurement*, *24*, 749–753.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, *84*, 289–297.

Kraemer, H. C. (1980). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika*, *44*, 461–472.

Landis, R. J., & Koch, G. G. (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.

Quenouille, M. H. (1949). Approximate tests of correlation in times series. *Journal of the Royal Statistical Society, B*, *11*, 68–84.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *XIX*, 321–325.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

Received 6 January 2006; revised version received 14 June 2006