Measurement and Research Department Reports 2001–3

About the Cluster Kappa Coefficient

T.M. Bechger

B.T. Hemker

G.K.J. Maris

Cito Arnhem

Measurement and Research Department Reports 2001–3

About the Cluster Kappa Coefficient

T.M. Bechger

B.T. Hemker

G.K.J. Maris

Citogroep
Arnhem, augustus 2001

Abstract

The cluster kappa was proposed by Schouten (1982) as a measure of chance-corrected rater agreement suitable for studies where objects are rated on a categorical scale by two or more judges. We discuss a way to calculate the cluster kappa which is suited even if ratings are missing. Further, we demonstrate how the sampling error of the cluster kappa may be estimated.

1 Introduction

This paper deals with a measure of chance-corrected rater agreement that is suitable for studies where objects are rated on a nominal or ordinal scale by two or more judges. This measure, which is called the *Cluster-kappa* (Cκ), was proposed by Schouten (1982; 1985, section 2.4). In this note we introduce a way to calculate the Cκ which is suited even when some ratings are missing by accident or on purpose. Furthermore, we show how the sampling error of the Cκ (and similar measures) may be estimated.

In Section 2, we introduce our notation. In Section 3, we introduce the Cκ. In Section 4, we discuss how the Cκ relates to Cohen's kappa (Cohen, 1960; 1968) and a number of published measures for rater agreement. In Section 5, we discuss how the sampling variance of the Cκ can be estimated. In Section 6, we analyze a real data set reported in Dunn (1989). Finally, in Section 7 we discuss and summarize our findings. Proofs of some statements in the text are presented in the Appendix where we also outline a computer program to calculate the Cκ for a given set of ratings.

2 Notation

We are interested in quantifying the agreement among multiple ratings of a random sample of N objects. The discussion is not limited to any particular kind of objects such as essays by foreign students, or biopsy slides. There are $R \geq 2$ raters, numbered from 1 to R, that have evaluated one or more of the objects. We use the generic words "raters" and "ratings" here to include observers, judges, diagnostic tests, etc. and their ratings/results. We

assume that raters work independently from one another, and that they are given instructions to assign an object to one of c categories. We regard each category as being labelled with an integer in the range from 1 to c.

We consider two mutually exclusive subsets or "clusters " of raters, Ω_1 and Ω_2, with $R^{(1)}$ and $R^{(2)}$ elements, respectively. Each cluster is formed by raters that are believed to originate from a common population of raters. At least the first cluster is non-empty, i.e., $\Omega_1 \neq \emptyset$. Although $R^{(1)}$ may be larger than 2, we shall consider agreement between two raters that are representative of the available raters. When $\Omega_2 = \emptyset$, both raters are randomly selected from Ω_1. Otherwise, the first rater is selected from Ω_1, and the second rater is selected from Ω_2.

Let x_{ij} denote the number of objects that were placed in category i by the first rater and in category j by the second rater. The letters i and j will be used throughout to denote two arbitrary categories. Let $x_{i+} \equiv \sum_{j=1}^{c} x_{ij}$, $x_{+i} \equiv \sum_{j=1}^{c} x_{ji}$, and $x_{++} \equiv \sum_{i=1}^{c} \sum_{j=1}^{c} x_{ij}$. The numbers x_{ij} are collected in an $c \times c$ *agreement matrix* $\mathbf{X} = (x_{ij})$. Let $p_{ij} \equiv \frac{x_{ij}}{N}$ denote the proportion of objects that are placed in category i by the first rater and in category j by the second rater. The proportions may be gathered in a matrix of proportions $\mathbf{P} = (p_{ij})$. Let the *marginal proportions* be defined as $p_{i+} \equiv \sum_{j=1}^{c} p_{ij}$, and $p_{+i} \equiv \sum_{j=1}^{c} p_{ji}$ for $i = 1, .., c$.

Let

$$N_p \equiv \begin{cases} R^{(1)}(R^{(1)} - 1) \text{ if } \Omega_2 = \emptyset \\ R^{(1)}R^{(2)} \text{ otherwise} \end{cases} . \tag{1}$$

When there is one cluster, N_p is equal to the number of ordered pairs of raters from Ω_1. When there are two clusters, N_p equals the number of combinations

of one rater from Ω_1 and another rater from Ω_2.

A different definition is necessary when some of the possible ratings are missing. Let v denote a generic object. Let $R_v^{(f)}$ denote the number of raters in cluster Ω_f that have evaluated object v. Finally, let $N_{p,v}$ equal the number of pairs of raters where both raters have rated object v.

$$N_{p,v} \equiv \begin{cases} \frac{1}{2} R_v^{(1)} (R_v^{(1)} - 1) \text{ if } \Omega_2 = \emptyset \\ R_v^{(1)} R_v^{(2)} \text{ otherwise} \end{cases} . \tag{2}$$

The rationale for the factor $\frac{1}{2}$ in this definition will be explained in Section 5.

3 The Cluster Kappa

Many indices for chance-corrected rater agreement are of the following type:

$$I = \frac{O - E}{M - E}, \tag{3}$$

where O denotes the observed degree of agreement, M denotes the maximum possible agreement, and E denotes the amount of chance agreement. If $I = 1$ there is perfect agreement ($O = M$), and if $I = 0$ the observed agreement is merely chance agreement ($O = E$). I becomes negative when the observed agreement is worse than chance.

We obtain the Cκ if we define O, M, and E as follows:

$$O = P_o \equiv \sum_{i=1}^{c} \sum_{j=1}^{c} \omega_{ij} p_{ij} \tag{4}$$

and

$$E = P_e = \sum_{i=1}^{c} \sum_{j=1}^{c} \omega_{ij} p_{i+} p_{+j}. \tag{5}$$

Observed agreement is defined as the proportion of objects that are placed in the same category by both raters, and chance agreement as the proportion agreement when measurements from the first rater are randomly matched to those of the second rater (Dunn, 1989, section 2.7; Fleiss, 1975). The maximum possible agreement is 1 which can only be obtained if $p_{i+} = p_{+i}$ for all i (Dunn, 1998, section 2.11).

The weights ω_{ij} are chosen on substantive grounds prior to the ratings to express the relative similarities among the categories. Following Yang en Chen (1978), we assume that $0 \leq \omega_{ij} \leq 1$, $\omega_{ii} = 1$, and $\omega_{ij} = \omega_{ji}$, for all $i, j = 1, ..., c$. This is not a serious restriction on the weights. If a researcher decides that the numbers d_{ij} $(i, j = 1, ..., c)$ express the difference between the categories, the weights can be calculated as

$$\omega_{ij} = 1 - \frac{d_{ij}}{\max_{i,j}(d_{ij})}. \tag{6}$$

It follows that

$$\mathrm{C}\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{7}$$

The maximum value of Cκ is 1. The minimum value of Cκ is -1 if $P_e = \frac{1}{2}$. Otherwise, the minimum values lies between 0 and -1. Thus, Cκ can assume values in the interval $[-1, 1]$. It is clear that the Cκ can not be calculated if $P_e = 1$, which happens if and only if both raters assign all objects to one category (see Appendix).

When $c > 2$, the value of the Cκ is dependent upon the weights. To be more specific, if we increase the value of w_{ij}, the value of Cκ increases if and

only if

$$\frac{p_{ij} + p_{ji}}{p_{i+}p_{j+} + p_{j+}p_{+i}} > 1 - C\kappa, \tag{8}$$

where $C\kappa$ is calculated using the original weights (Schouten, 1985, section 1.3; 1986, p. 455). Schouten (*1985,* section 3.3 and 1986, section 4) demonstrates how this results is used to find categories that are relatively hard to distinguish (see Section 6 for an application).

How do we calculate the agreement matrix ? Suppose that we consider agreement among $R^{(1)} \leq R^{(1)}$ raters in a single cluster Ω_1. For each object, we may randomly draw (with replacement) n pairs of ratings out of the available ratings and count the number of pairs in each cell of $\mathbf{X}$, which is then used to calculate $C\kappa$. In this situation, Schouten (1985) appropriately refers to $C\kappa$ as the *intra-cluster kappa.* Both raters represent the average of the raters in Ω_1 and $\mathbf{X}$ will be symmetric. If $n \rightarrow \infty$, ratings by each of the N_p ordered pairs of raters will be chosen $N_p^{-1}n$ times. This means that if, for each object, we consider the ratings of all unordered pairs of raters, we obtain the asymptotic results without sampling. This will usually be preferable unless $R^{(1)}$ is very large. Note that in practice only $\frac{1}{2}N_p$ pairs of ratings need to be considered (see Appendix).

The situation is different when we distinguish two clusters of raters, for example, $R^{(1)}$ male and $R^{(2)}$ female raters. In this situation, we would first randomly choose a male rater, and then a female rater so that the first rater is an average male rater, and the second rater is an average female rater. Similar to the one-cluster case, we get the same result if we use all NN_p combinations of a rating by a male rater and a rating by a female rater to construct $\mathbf{X}$ and calculate $C\kappa$. In the two-cluster case, Schouten names the

Cκ an *inter-cluster kappa*. Now, $\mathbf{X}$ need not be symmetric. Following the analysis, we could consider agreement among raters within each cluster.

When some ratings are missing, only those pairs are counted for which both ratings are available and we assume that the average ratings are not systematically different from the average ratings that would have been obtained with complete data. For example, when only two out of $R^{(1)}$ judgements have been observed per object, we assume that the data set is actually a random draw from the complete data with either one cluster or with two clusters. We have recently applied the Cκ in two studies of this kind. In both studies, only two out of thirty raters evaluated the performance of each of a group of students. To be more specific, for each student anew the two raters were chosen randomly from the available raters. In the first study, there was no reason to assume that the raters originated from different populations and we calculated the intra-cluster kappa. In the second study, the first rating was always done by someone who knew the student, while the second rating was done by someone who did not know the student. In this situation, we calculated the inter-cluster kappa; Ω_1 consisted of raters that are familiar with the students, and Ω_2 of raters that are not familiar with the student.

In general, to apply the Cκ in a situation where ratings are missing, the design of the study should be such that, for each object, ratings within a cluster of raters are missing completely at random. Such designs may be made automatically using a random number generator.

4 Relations to Other Measures of Rater Agreement

If no ratings are missing and $R^{(1)} = R^{(2)} = 1$, the Cκ is equal to Cohen's kappa (Cohen, 1960;1968). Cohen devised his measure to improve upon a measure that was proposed by Scott (1955), which equals the Cκ if $\Omega_2 = \emptyset$ and $R^{(1)} = 2$. Basically, all measures that are mentioned in this paper, are based upon Cohen's kappa. The Cκ, in particular, is equivalent to Cohen's kappa applied to ratings of two raters that are representative of one or two kinds of raters.

The Cκ is also related to the *Weighted Mean kappa*

$$\text{WM}\kappa = \frac{\sum_{g,h}(1 - P_e^{(g,h)})C\kappa^{(g,h)}}{\sum_{g,h}(1 - P_e^{(g,h)})}, \tag{9}$$

which was proposed by Hubert (1977; Schouten, 1985). The symbol $\sum_{g,h}$ indicates that we summate across N_p rater pairs. Superscript (g, h) indicates that a statistic is calculated using only ratings by raters g and h. The WMκ is equal to the inter-cluster kappa. It equals the intra-cluster kappa when $R^{(1)} \rightarrow \infty$, or when $P_e^{(g,h)} = P_e$ for all pairs of raters in Ω_1 (Schouten, 1985, pp. 52-53).

When $P_e^{(g,h)} = P_e$ for all pairs of raters (either in the one – or in the two-cluster case), the WMκ is equal to

$$\overline{C\kappa} = \frac{\sum_{g,h} C\kappa^{(g,h)}}{N_p} \tag{10}$$

which was proposed as a measure of agreement by Light (1971). It follows that WMκ =Cκ = $\overline{C\kappa}$ if $P_e^{(g,h)} = P_e$ for all pairs of raters.

Finally, we note that the intra-cluster kappa is equal to a measure proposed by Fleiss (1971). Hence, the reader is referred to Fleiss (1971) for an alternative scheme to calculate the intra-cluster kappa.

5 The Sampling Distribution of the Cκ

Let Cκ denote the population value, and C$\hat{\kappa}$ the value found in a sample of ratings of N objects. To determine the distribution of C$\hat{\kappa}$ we use the *non-parametric bootstrap method* (Efron & Tibshirani, 1998). This method entails random sampling of the ratings of N objects from the data, without replacement. For each bootstrap sample the C$\hat{\kappa}$ is calculated. The empirical distribution over the bootstrap samples is a consistent estimate of the distribution of the C$\hat{\kappa}$ and it may be used to construct a confidence interval around C$\hat{\kappa}$ (Efron & Tibshirani, 1998, chapters 12-14, and 22).

To estimate the variance of C$\hat{\kappa}$, the bootstrap will usually require between 50 and 200 samples. To monitor the convergence of the bootstrap we like to plot the estimates against the number of bootstrap samples. In Figure 1 the scale was fixed after 200 samples to see if the estimates were within $\pm 3\,|\mathrm{C}\hat{\kappa}/1000|$ of the estimated Cκ across samples. Figure 1 also shows the estimated sampling density. Various data sets published in Schouten (1985) and Dunn (1989) were analyzed (see Section 6) and it was found that the sampling distribution looks like a normal distribution, except when N is very small. This is also what one expects from the multivariate central limit theorem (Rao, 1973).

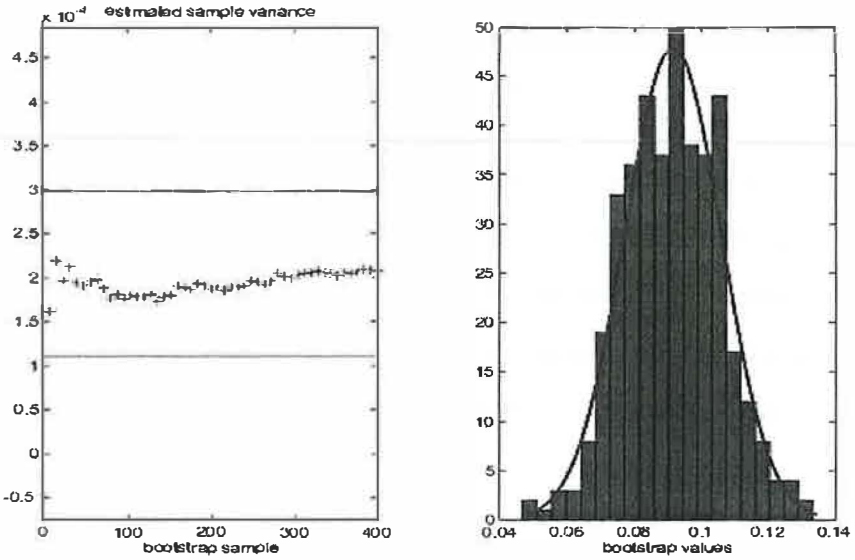


Figure 1: The first figure is a plot of the estimated variance of $\hat{C\kappa}$ against the number of bootstrap samples. The second figure shows the empirical distribution of $\hat{C\kappa}$. Further details of this analysis are in Section 6.

To test the hypothesis that $\hat{C\kappa} = 0$, we can calculate the statistic

$$z_0 = \frac{\hat{C\kappa}}{\sqrt{\sigma_0^2(\hat{C\kappa})}}, \tag{11}$$

where $\sigma_0^2(\hat{C\kappa})$ denotes the variance of $\hat{C\kappa}$ when $C\kappa = 0$. When $\hat{C\kappa} \sim \mathcal{N}(0, \sigma_0^2(\hat{C\kappa}))$, $z_0 \sim \mathcal{N}(0,1)$ and the probability of finding $\hat{C\kappa}$, given $C\kappa = 0$, can be found in tables of the standard normal distribution.

We will now derive an approximate expression for $\sigma_0^2(\hat{C\kappa})$; the variance if $C\kappa = 0$, and $P_e^{(g,h)} = P_e$ for all pairs of raters. If N is large so that the

marginal probabilities are known

$$\mathrm{C}\hat{\kappa} = \frac{\sum_{v=1}^{N} \mathrm{C}\hat{\kappa}_v}{N}, \tag{12}$$

where $\mathrm{C}\hat{\kappa}_v$ denotes the $\mathrm{C}\kappa$ calculated from the agreement matrix of object v, based upon $N_{p,v}$ pairs of raters. The right side of (12) is no more than an alternative way to calculate $\overline{C\kappa}$ (see Equation 10). If objects are rated independently of one another, (12) implies that $\mathrm{C}\hat{\kappa}$ is the mean of N independently, and identically distributed $\mathrm{C}\hat{\kappa}_v$'s so that

$$\sigma_0^2(\mathrm{C}\hat{\kappa}) = \frac{1}{N^2} \sum_{v=1}^{N} \sigma_0^2(\mathrm{C}\hat{\kappa}_v; N_{p,v}), \tag{13}$$

where $\sigma_0^2(\mathrm{C}\hat{\kappa}_v; N_{p,v})$ denotes the sampling variance of $\mathrm{C}\hat{\kappa}_v$ given that $\mathrm{C}\kappa_v = 0$. With the Delta method, Fleiss, Cohen & Everitt (1969) found that

$$\sigma_0^2(\mathrm{C}\hat{\kappa}_v; N_{p,v}) \approx \frac{1}{N_{p,v}(1-P_e)^2} \sum_{i=1}^{c} \sum_{j=1}^{c} p_{i+} p_{+j} (w_{ij} - \bar{\omega}_{i+} - \bar{\omega}_{+j})^2 - (P_e)^2, \tag{14}$$

where

$$\bar{\omega}_{i+} = \sum_{j=1}^{c} \omega_{ij} p_{+j}, \quad \bar{\omega}_{+j} = \sum_{i=1}^{c} \omega_{ij} p_{i+}. \tag{15}$$

When $\Omega_2 = \emptyset$, the factor $\frac{1}{2}$ in $N_{p,v}$ (see Equation 2) is appropriate since pair $\{i, j\}$ contributes exactly the same information as pair $\{j, i\}$. We obtain an approximate expression for $\sigma_0^2(\mathrm{C}\hat{\kappa})$ if we substitute (14) in (13). Note that this approximation also applies to the other measures mentioned in Section 4.

Schouten (1985, section 4.3) and Fleiss et al. (1979) also derived $\sigma_0^2(\mathrm{C}\hat{\kappa}_v; N_{p,v})$ but only for the intra-cluster kappa. Our simulations and simulations by

Fleiss, Nee & Landis (1979) suggest that the distribution of z_0 approximates a standard normal distribution when N is over 25, and the marginal proportions are not very different.

6 An Application

Dunn (1989, table 7.11) reports ratings of 181 conference abstracts by eight independent referees. Apart from missing ratings there were four categories; (A) Strongly recommended as a spoken contribution to the conference, (B) Recommended as a spoken contribution ,(C) Recommended for acceptance as a short communication on a poster, and (D) Recommended for rejection. There were 139 (about 9.6% of the data) missing ratings about 56% of which were due to the second and third rater. We do not know anything about the raters and we assume that they constitute a single cluster.

With a single cluster of raters, we found that $C\hat{\kappa} = 0.0931$, $\sigma_0^2(C\hat{\kappa}) = (0.01099)^2$ and $z_0 = 8.5$. Although we reject the hypothesis that $C\kappa = 0$, the observed value is quite close to zero and we feel confident to conclude that there is little agreement among the referees. The bootstrap standard error was found to be 0.0134 and a normal theory 90% confidence interval for $C\hat{\kappa}$ is $[0.07 - 0.12]$ (see Figure 1). Both the $\overline{C\kappa}$ (Equation 10) and the WMκ (Equation 9) were estimated to be 0.1. The abstracts that were not rated by both raters were simply ignored in the calculation.

If we look at the ratio's of observed versus expected confusion among categories (i.e. Equation 8) we find that $C\kappa$ can be increased by combining categories A with B, and C with D. The new categories are: (A*) recommend

as a spoken contribution, and (B*) reject as a spoken contribution. It appears that the referees are more unanimous in their recommendation to accept an abstract or not; the $C\hat{\kappa}$ is now 0.21 and the 90% confidence interval is [0.17 – 0.26]. The confidence intervals reveal that the improvement in agreement is statistically significantly. We may also employ the bootstrap method to estimate the sampling variance of the difference between the $C\hat{\kappa}^{(1-2,3-4)}$ with the categories combined and $C\hat{\kappa}^{(1,2,3,4)}$ without the categories combined. In the present application, we find that

$$z = \frac{C\hat{\kappa}^{(1-2,3-4)} - C\hat{\kappa}^{(1,2,3,4)}}{\sqrt{\sigma^2(C\hat{\kappa}^{(1-2,3-4)} - C\hat{\kappa}^{(1,2,3,4)})}} \frac{0.21 - 0.093}{0.0165} = 7.25. \qquad (16)$$

As a next step, it is often useful to look for those objects where two categories are confused relatively often. These objects may be discussed with the raters in order to gain an understanding of the reason for such confusion and to improve future ratings. To this aim, we calculated for each abstract the product of the percentages of ratings in category 1 and category 2. The closer this value is to $\frac{1}{4}$ the more the categories are confused in the rating of an abstract. This procedure points towards the abstracts which where numbered 171 to 177 in the data.[1]

Considering the marginal distributions of the raters it is doubtful whether they constitute a single cluster. To facilitate the recognition of differences in marginal proportions we use pie-charts as in Figure 2. In Figure 2, we immediately spot three clusters: raters 1, 2 and 3, who reject about 63% of the abstracts, raters 2, 4, and 5 who reject slightly about half the abstracts,

[1] It is unclear why these abstracts have succesive numbers. It is possible that the abstracts have somehow been ordered.

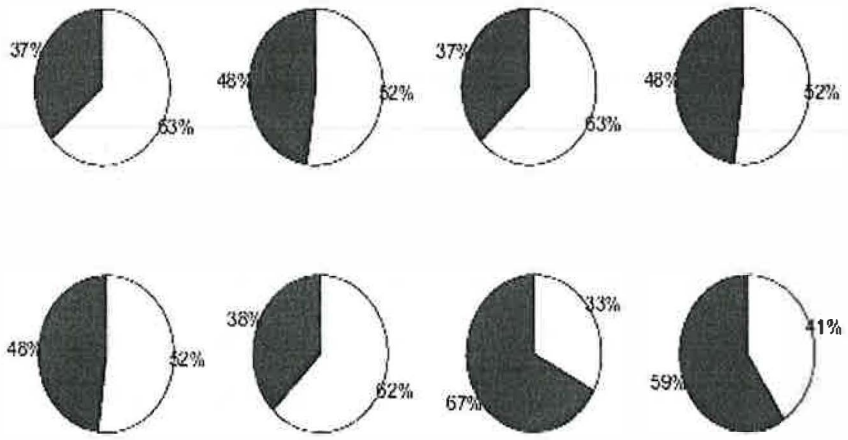


Figure 2: Pie charts of marginal probablities of the eight referees in the data reported by Dunn (1989, table 7.11).

and 7 and 8 who reject most abstracts that they were asked to evaluate. To improve the analysis we consider an inter-cluster kappa with two clusters $\Omega_1 = \{1, ..., 6\}$, and $\Omega_2 = \{7, 8\}$. The intercluster kappa was estimated to equal $C\hat{\kappa} = 0.169$, $\sigma_0^2(C\hat{\kappa}) = (0.022)^2$ with a confidence interval of $[0.11 - 0.23]$. The agreement between the two clusters can be summarized in

$$\mathbf{P} = \begin{pmatrix} 0.25 & 0.30 \\ 0.12 & 0.32 \end{pmatrix}. \tag{17}$$

These proportions clearly show that members of the second cluster are inclined to reject an abstract even when a rater from the first cluster has accepted the abstract as a spoken contribution to the conference. The intra-

cluster kappa's are as follows

$$\Omega_1 : \{1,2,3,4,5,6\}, \mathrm{C}\hat{\kappa} = 0.249, \sigma_0^2(C\hat{\kappa}) = (0.022)^2, [0.20 - 0.30] \quad (18)$$

$$\Omega_2 : \{7,8\}, \mathrm{C}\hat{\kappa} = 0.213, \sigma_0^2(C\hat{\kappa}) = (0.079)^2, [0.08 - 0.35] \quad (19)$$

In the absence of some 'golden standard ' we are unable to say whether raters in Ω_1 are more accurate than raters in Ω_2. It may be worthwhile to invite referees from both clusters and have them settle their differences.

Although it is hardly opportune in the present example, we may also consider clustering on the basis of agreement using some algorithm for cluster analysis (see Schouten, 1985). In general, it is quite likely that we find the same clusters that we find by looking at the marginal proportions because of the influence of the marginal proportions on the Cκ. To eliminate the influence of the marginal proportions we could divide the cluster kappa's by the maximum value over the set of agreement matrices with the observed marginal proportions (Dunn, 1998, section 2.11). Calculating this maximum is straightforward when $w_{ij} = 0$ if $i \neq j$. In that case, Cκ is maximal if $P_o = \sum_{i=1}^{c} \min(p_{i+}, p_{+i})$. In general, however, when $w_{ij} > 0$ for some $i \neq j$, there appears to be no other option than to enumerate all possible agreement matrices with the observed marginal proportions (see Appendix). Developing a general and fast algorithm to find the maximum value of Cκ is a topic for future research.

7 Conclusion

In this paper we have discussed an alternative way to calculate the Cκ, which was cut out to handle missing ratings. Furthermore, we have provided

ways to calculate the sampling variance of the Cκ that were not previously available.

A weakness that the Cκ shares with all measures based upon Cohen's kappa is that its value is difficult to interpret.[2] Suppose that there are N_a rater pairs that disagree on one or more objects. It can be shown (see Appendix) that

$$P_o = 1 - \left(\frac{N_a \sum_{i=1}^{c} \sum_{j=1, j \neq i}^{c} (1 - \omega_{ij}) \bar{x}_{ij}^{(a)}}{N_p \bar{x}_{++}} \right), \tag{20}$$

where $\bar{x}_{ij}^{(a)}$ denotes the mean number of objects in cell i, j across the N_a pairs of differing raters, $\bar{x}_{++}$ equals the mean number of objects graded by rater pairs, and $(1 - \omega_{ij})$ denotes the difference between category i and category j. The second term on the right side of Equation 20 is positive unless $\bar{x}_{ij}^{(a)} = 0$. Hence, the value of P_o is dependent upon the number of raters that disagree with one another, and the number of objects that they disagree about. The Cκ can only be unity if all rater pairs are unanimous. When C$\kappa < 1$ it can easily be adapted to provide more detailed information about the ratings. We may look for homogeneous clusters of raters, compare different raters to a standard, or investigate whether some categories are more often confused than others. All this is discussed in detail by Fleiss (1981), and Schouten (1985).

We used the bootstrap method because it is easy to implement in a computer program and easily adapted to serve in non-standard situations.[3] We mention three such situations. First, suppose that raters have evaluated the

[2] Some rules-of-thumb for the interpretation of values between 0 and 1 are given by Landis & Koch (1977).

[3] Note that Schouten (1985; 1986) recommended the jackknife procedure (Tukey, 1958).

same objects twice, using different scales, and we wish to know on which scale raters agree most. To get the answer, we may calculate C$\hat{\kappa}$ on both occasions, giving us C$\hat{\kappa}^{(1)}$ and C$\hat{\kappa}^{(2)}$, and investigate the significance of the difference C$\hat{\kappa}^{(1)}$–C$\hat{\kappa}^{(2)}$. To this aim, we may use the statistic

$$z = \frac{C\hat{\kappa}^{(1)} - C\hat{\kappa}^{(2)}}{\sigma^2(C\hat{\kappa}^{(1)} - C\hat{\kappa}^{(2)})}, \tag{21}$$

where $\sigma^2(C\hat{\kappa}^{(1)} - C\hat{\kappa}^{(2)})$ denotes the sampling variance of the difference. To estimate $\sigma^2(C\hat{\kappa}^{(1)} - C\hat{\kappa}^{(2)})$ we can calculate $C\hat{\kappa}^{(1)} - C\hat{\kappa}^{(2)}$ for different bootstrap samples of the objects and estimate its variance. Second, we could consider calculating the Cκ for different clusters or "strata " of objects, combine those estimates to estimate a common underlying Cκ, and use the bootstrap method to calculate the variance of the estimate (see Barlow, Lai & Azen, 1991; Lui & Kelly, 1999). The work of Barlow, et al. (1991) suggests that failure to take account of different groups of objects may produce a misleading estimate of the Cκ. Finally, suppose that, per subject, we calculate the median over rater pairs. The reason could be that the median is more robust against outliers among the rater pairs, and the resulting Cκ may therefore be called *the robust Cκ*. The bootstrap method can easily be used to estimate the sampling variance of the robust Cκ.

Although we acknowledge the advantage of methods based on latent trait models (e.g., Klauer & Batchelder, 1996), we note that estimation may break down when there are many missing ratings. We believe that the difficulty in interpreting intermediate values of the C$\hat{\kappa}$ is counterbalanced by its wide

Although we found the jackknife to perform well, the bootstrap is to be preferred on theoretical grounds because the Cκ is a non-linear statistic (see Efron & Tibshirani, 1998, §11.5).

range of applications and the possibilities it offers to investigate the reasons for low agreement.

8 Appendix

8.1 $P_e = 1$ If and Only If Raters Use Only One Category

Assume that $0 \leq \omega_{ij} \leq 1$, $\omega_{ii} = 1$ and $\omega_{ij} = \omega_{ji}$, for all i, j. We also assume that $w_{ij} > 0$, for some i, j.

$$P_e = 1 \Leftrightarrow 1 - P_e = 0 \Leftrightarrow \tag{22}$$

$$\sum_{i=1}^{c}\sum_{j=1}^{c} p_{i+}p_{+j} - \sum_{i=1}^{c}\sum_{j=1}^{c} \omega_{ij}p_{i+}p_{+j} \tag{23}$$

$$= \sum_{i=1}^{c}\left[\sum_{j\neq i}^{c}(1-\omega_{ij})p_{i+}p_{+j}\right] = 0 \tag{24}$$

Each term within brackets is positive and they must all be equal to 0. Suppose that $p_{h+} > 0$, where h is an arbitrary category. If $i = h$,

$$\left[\sum_{j\neq h}^{c}(1-\omega_{hj})p_{h+}p_{+j}\right] = 0 \Leftrightarrow \tag{25}$$

$$\sum_{j\neq h}^{c}(1-\omega_{hj})p_{+j} = 0 \Leftrightarrow p_{+j} = 0,\ \forall j \neq h \tag{26}$$

Since $\sum_{i=1}^{c} p_{+i} = 1$, this implies that $p_{+h} = 1$. In the same manner we find that $p_{h+} = 1$. Hence, if $P_e = 1$ all raters use only one category. It is easy to see that the opposite is also true, i.e., $p_{h+}p_{+h} = 1 \Rightarrow P_e = 1$.

8.2 The Derivation of Expression 20

Let $\sum_{g,h}$ denote summation across rater pairs; $\sum_{a}$ denotes summation over N_a pairs of raters that are not unanimous and $\sum_{-a}$ denotes summation over

all unanimous pairs of raters. Let $x_{ij}^{(g,h)}$ denote the number of objects placed in cell (i,j) of $\mathbf{X}$ by raters g and h. We find that

$$\begin{aligned}
P_o &= \sum_{i=1}^{c}\sum_{j=1}^{c}\omega_{ij}\frac{\sum_{g,h}x_{ij}^{(g,h)}}{\sum_{g,h}x_{++}^{(g,h)}} \\
&= \sum_{i=1}^{c}\frac{\sum_{g,h}x_{ii}^{(g,h)}}{\sum_{g,h}x_{++}^{(g,h)}} + \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\omega_{ij}\frac{\sum_{a}x_{ij}^{(a)}}{\sum_{g,h}x_{++}^{(a)}} \\
&= \sum_{i=1}^{c}\frac{\bar{x}_{ii}}{\bar{x}_{++}} + \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\omega_{ij}\frac{N_a\bar{x}_{ij}^{(a)}}{N_p\bar{x}_{++}} \\
&= \sum_{i=1}^{c}\frac{\bar{x}_{ii}}{\bar{x}_{++}} + \frac{N_a}{N_p\bar{x}_{++}}\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\omega_{ij}\bar{x}_{ij}^{(a)} \\
&= \frac{1}{\bar{x}_{++}}\left(\sum_{i=1}^{c}\bar{x}_{ii} + \frac{N_a}{N_p}\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\omega_{ij}\bar{x}_{ij}^{(a)}\right)
\end{aligned}$$

To simplify the term within brackets we write:

$$\begin{aligned}
\sum_{i=1}^{c}\bar{x}_{ii} &= \sum_{i=1}^{c}\frac{1}{N_p}\sum_{g,h}x_{ii} \\
&= \frac{1}{N_p}\sum_{g,h}\sum_{i=1}^{c}x_{ii} \\
&= \frac{1}{N_p}\left(\sum_{-a}\left(\sum_{i=1}^{c}x_{ii}^{(g,h)}\right) + \sum_{a}\sum_{i=1}^{c}x_{ii}^{(a)}\right) \\
&= \frac{1}{N_p}\left(\sum_{-a}x_{++}^{(g,h)} + \sum_{a}x_{++}^{(pk)} - \sum_{a}\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}x_{ij}^{(a)}\right) \\
&= \frac{1}{N_p}\left(\sum_{g,h}x_{++}^{(g,h)} - \sum_{a}\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}x_{ij}^{(a)}\right) \\
&= \bar{x}_{++} - \frac{\sum_a\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}x_{ij}^{(a)}}{N_p}
\end{aligned} \tag{27}$$

So that

$$P_o = \frac{1}{\bar{x}_{++}}\left(\bar{x}_{++} - \frac{\sum_a \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c} x_{ij}^{(a)}}{N_p} + \frac{N_a}{N_p}\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\omega_{ij}\bar{x}_{ij}^{(a)}\right) \tag{28}$$

$$= 1 - \left(\frac{\sum_a \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c} x_{ij}^{(pk)} - \sum_a \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\omega_{ij}x_{ij}^{(a)}}{N_p\bar{x}_{++}}\right)$$

$$= 1 - \left(\frac{\sum_a \left(\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c} x_{ij}^{(a)} - \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\omega_{ij}x_{ij}^{(a)}\right)}{N_p\bar{x}_{++}}\right)$$

$$= 1 - \left(\frac{\sum_a \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}(x_{ij}^{(a)} - \omega_{ij}x_{ij}^{(a)})}{N_p\bar{x}_{++}}\right)$$

$$= 1 - \left(\frac{\sum_a \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}(1-\omega_{ij})x_{ij}^{(a)}}{N_p\bar{x}_{++}}\right)$$

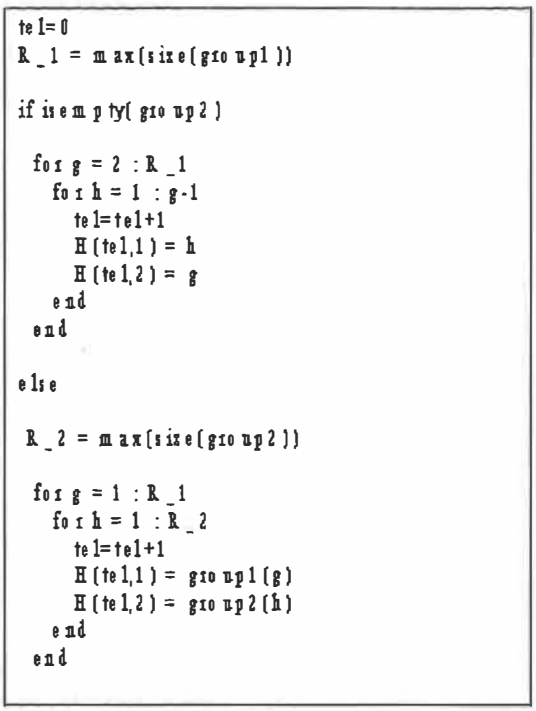
$$= 1 - \left(\frac{N_a \sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}(1-\omega_{ij})\bar{x}_{ij}^{(a)}}{N_p\bar{x}_{++}}\right)$$

We have now arrived at Equation 20. With complete data, this formula applies to all measures mentioned in Section 4 of this paper.

8.3 An Algorithm for the Calculation of the Cκ

Let the categories in the data be coded with integers from 1 to c. The raters are arbitrarily numbered from 1 to R. The first step in the actual calculation of the Cκ is the construction of a matrix $\mathbf{H}$ whose rows are the numbers of the rater pairs involved in the calculation (see Figure 3). When $\Omega_2 = \emptyset$, the rows of $\mathbf{H}$ are $\frac{1}{2}N_p$ combinations of two out of $R^{(1)}$. It is not necessary to consider all N_p ordered pairs in the calculation. When $\Omega_2 \neq \emptyset$, the rows of $\mathbf{H}$ are N_p combinations of two raters; one from Ω_1, and one from Ω_2.

Following the construction of $\mathbf{H}$, ratings are summed over objects and



```
tel=0
R_1 = max(size(group1))

if isempty(group2)

 for g = 2 : R_1
   for h = 1 : g-1
     tel=tel+1
     H(tel,1) = h
     H(tel,2) = g
   end
 end

else

 R_2 = max(size(group2))

 for g = 1 : R_1
   for h = 1 : R_2
     tel=tel+1
     H(tel,1) = group1(g)
     H(tel,2) = group2(h)
   end
 end
```

Figure 3: Pseudo-code for a computer program that constructs the matrix **H**. The numbers of the raters in each cluster are in the arrays group1 and group2 of length R_1 and R_2, respectively.

```
lambda = 0
N_p = max(size(H))
  for v = 1 : N
     N_pv = 0
     for g = 1 : N_p
       if data(v,H(g,1))~=mis_code & data(v,H(g,2))~=mis_code
         N_pv = N_pv + 1
         x(data(v,H(g,1)),data(v,H(g,2))) = x(data(v,H(g,1)),data(v,H(g,2))) + 1
       end
     end
     if N_pv > 0
      lambda  = lambda + 1/N_pv;
     end
   end

  P = x./sum(sum(x))

if isempty(group2)
   P = (0.5)*(P + P')
end
```

Figure 4: The matrix **H** and the data are used to calculate the mean agreement matrix with elements x(i,j). mis_code is the code that is used for missing data.

over pairs of raters (see Figure 4). The resulting agreement matrix is used to calculate the proportions. However, when there is one group of raters, the matrix of proportions must first be made symmetric because we have only considered $\frac{1}{2}N_p$ unordered pairs. Once we have the proportions, Cκ is calculated via Formulae 5, 4 and 7.

Write $\sigma_0^2(C\hat{\kappa})$ as $\frac{A\lambda}{N^2}$, where

$$A = \frac{1}{(1-P_e)^2}\sum_{i=1}^{c}\sum_{j=1}^{c} p_{i+}p_{+j}(w_{ij} - \bar{\omega}_{i+} - \bar{\omega}_{+j})^2 - (P_e)^2 \qquad (29)$$

and

$$\lambda = \sum_{v=1}^{N} \frac{1}{N_{p,v}} \tag{30}$$

The symbols are as defined above. The term λ is calculated when we construct $\mathbf{X}$, as can be seen in Figure 4. The term A is based upon the matrix of proportion $\mathbf{P}$. Of course, when we use the algorithm to calculate the bootstrap we have no need to calculate $\sigma_0^2(C\hat{\kappa})$.

8.4 Calculating the Maximum Value of Cκ Given the Marginal Proportions

Suppose we observe an agreement table. There are many tables with the observed marginal proportions; together they form a so-called *isomarginal family*. We need to find the member of this family that maximized the value of P_o.

We will first, very briefly, describe how we approached the problem of finding the maximum value of Cκ and then give an example where it fails. Let Ω_X denote the set of $c \times c$ matrices with the same row and column sums as the observed agreement matrix $\mathbf{X}$. Let $\mathbf{T} = (t_{ij})$ denote a generic member of Ω_X. Our task is to find a matrix $\mathbf{T}$ such that

$$P_o(t_{ij})x_{++} = w_{11}t_{11} + w_{12}t_{12} + ... + w_{cc}t_{cc}, \tag{31}$$

is maximized, where w_{ij} denotes an entry in the weight matrix $\mathbf{w}$. We assume that $\mathbf{w}$ has the following properties:

1. $\mathbf{w}$ is symmetric,

2. The diagonal elements of $\mathbf{w}$ are 1 and the off-diagonal elements are smaller than 1.

3. All entries in any given row or column are different. This implies that there is always a unique entry in each row of column that is associated with the largest weight.

We started from the following idea. We can write $\mathbf{T}$ as the sum of a symmetric matrix $\mathbf{S}^{(1)} = (s_{ij}^{(1)})$ and a rest-matrix $\mathbf{R}^{(1)} = \mathbf{T} - \mathbf{S}$; $\mathbf{R}^{(1)} = (r_{ij}^{(1)})$. If we first choose the maximum among the symmetric matrices and then the maximum among the rest matrices, we find the maximum $\mathbf{T}_{\max}$ over Ω_X by adding the two.

The matrix $\mathbf{S}^{(1)}$ is not unique but it seems reasonable to demand that the row (and column sums) of $\mathbf{S}^{(1)}$ are given by $\min(x_{i+}, x_{+i})$ for $i = 1, .., c$. Since $\mathbf{S}^{(1)}$ is symmetric it is clear that $P_o(s_{ij}^{(1)})s_{++}^{(1)}$ is maximized when it is diagonal, that is when

$$s_{ii}^{(1)} = \min(x_{i+}, x_{+i}),\ (i = 1, .., c) \tag{32}$$

and $s_{ij}^{(1)} = 0$ for $i \neq j$. The row – and column sums of $\mathbf{R}^{(1)}$ are $r_{i+}^{(1)} = x_{i+} - s_{ii}^{(1)}$, and $r_{+i}^{(1)} = x_{+i} - s_{ii}^{(1)}$, respectively. The diagonal elements of $\mathbf{R}^{(1)}$ are zero and the off-diagonal entries of $\mathbf{R}^{(1)}$ are given by t_{ij}.

After this first step, we continue to decompose $\mathbf{R}^{(1)}$ as the sum of a symmetric matrix $\mathbf{S}^{(2)}$ with $s_{ii}^{(2)} = \min(r_{i+}^{(1)}, r_{+i}^{(1)})$ (for $i = 1, .., c$) and a rest matrix $\mathbf{R}^{(2)}$, etc. We continue to decompose the rest matrices until all remaining elements of $\mathbf{T}$ are determined or until there are no non-zero weights associated with the remaining unknown entries in $\mathbf{T}$.

Before we continue to discuss further details we try the algorithm on a small example. Suppose

$$\mathbf{X} = \begin{pmatrix} 4 & 3 & 2 \\ 1 & 7 & 0 \\ 5 & 2 & 1 \end{pmatrix}, \mathbf{w} = \begin{pmatrix} 1 & 0.9 & 0.8 \\ 0.9 & 1 & 0.1 \\ 0.8 & 0.1 & 1 \end{pmatrix} \Rightarrow C\kappa = 0.3947 \tag{33}$$

The column sums are $(9, 8, 8)^t$ and the row sums are $(10, 12, 3)$. We decompose $\mathbf{T}$ as

$$\mathbf{T} = \mathbf{S} + \mathbf{R} \tag{34}$$

$$= \begin{pmatrix} 9 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 3 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ t_{31} & t_{32} & 0 \end{pmatrix} \tag{35}$$

The column sums of $\mathbf{R}$ are $(0, 0, 5)^t$ and the row sums are $(1, 4, 0)$. It follows that $t_{31} = 1$ and $t_{32} = 4$. Thus we find

$$\mathbf{T} = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 8 & 0 \\ 1 & 4 & 3 \end{pmatrix}, \Rightarrow C\kappa = 0.3540 \tag{36}$$

Quite surprisingly, the Cκ of $\mathbf{T}$ is actually smaller than the Cκ of $\mathbf{X}$. If we had decomposed

$$\mathbf{T}^* = \mathbf{S}^* + \mathbf{R}^* \tag{37}$$

$$= \begin{pmatrix} 4 & 1 & 2 \\ 1 & 7 & 0 \\ 2 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 3 & 2 & 0 \end{pmatrix} \tag{38}$$

we would have found $C\kappa = 0.5614$.

This example shows that it is wrong to consider only symmetric matrices with $s_{ii} = \min(x_{i+}, x_{+i})$. Each choice for a particular symmetric matrix generates a set of rest matrices where there is a matrix which gives the maximum value of $\mathbf{T}$. However, we have to walk through the set of all symmetric matrices with $s_{ii} \leq \min(x_{i+}, x_{+i})$ to find the maximum across the isomarginal family of $\mathbf{X}$.

We do not recognize the present problem as a instance of any standard problem and we hope that readers might provide suggestions for its solution.

9 References

Barlow, W., Lay, M-Y, & Azen, S. P. (1991). A comparison of methods for calculating a stratified kappa. *Statistics in Medicine*, *10*, 1465-1472.

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.

Dunn, G. (1989). *Design and analysis of reliability studies: The statistical evaluation of measurement errors*. New-York: Oxford University Press.

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. Chapman & Hall/CRC: London.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *30*, 469-479.

Fleiss, J.L. (1975). Measuring agreement between two judges on the presence and absence of a trait. *Biometrics*, *31*, 651-659.

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

Fleiss, J.L., Nee, J.C.M., & Landis, J.R. (1979). Large sample variance of kappa in the case of different sets of raters.*Psychological Bulletin*, *86*, 974-977.

Klauer, K.C., & Batchelder, W. H. (1996). Structural analysis of subjective categorical data. *Psychometrika*, *61*, 199-240.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

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

Lui, K.J., & Kelly, C. (1999). A note on interval estimation of kappa in a series of 2×2 tables. *Statistics in Medicine*, *18*, 2041-2049.

Schouten, H. J. A. (1982). Measuring pairwise agreement among many observers II: Some improvements and additions. *Biometrical Journal*, *24*, 431-435.

Schouten, H. J. A. (1985). *Statistical measurement of interobserver agreement: Analysis of agreements and disagreements between observers.* Unpublished Doctoral Dissertation: Erasmus University Rotterdam.

Schouten, H. J.A. (1986). Nominal scale agreement among observers. *Psychometrika*, *51*, 453-466.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New-York: Wiley.

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

Yang, G.L., & Chen, M.K.(1978). A note on weighted kappa. *Socio Economic Planning Sciences*, *12*, 293-294.