A review of agreement measure as a subset of association measure between raters

A. O. Adejumo, C. Heumann and H. Toutenburg
Department of Statistics
Ludwig-Maximilians-Universität, München
Ludwigstr. 33, D-80539 Munich,
Germany

26th May 2004

Abstract

Agreement can be regarded as a special case of association and not the other way round. Virtually in all life or social science researches, subjects are being classified into categories by raters, interviewers or observers and both association and agreement measures can be obtained from the results of this researchers. The distinction between association and agreement for a given data is that, for two responses to be perfectly associated we require that we can predict the category of one response from the category of the other response, while for two response to agree, they must fall into the identical category. Which hence mean, once there is agreement between the two responses, association has already exist, however, strong association may exist between the two responses without any strong agreement. Many approaches have been proposed by various authors for measuring each of these measures. In this work, we present some up till date development on these measures statistics.

keywords: Agreement, association, raters, kappa, loglinear, latent-class.

1 Introduction

Measures of association reflect the strength of the predictable relationship between the ratings of the two observers or raters. Measures of agreement pertain to the extent to which they classify a given subject identically into the same category. As such, agreement is a special case of association. If agreement exists between two observers, association also will definitely exist, but there can be strong association without strong agreement. For example, if in an ordinal scale, rater 1 consistently rates subjects one level higher than rater 2, then the strength of agreement is weak even though the association is strong. In social and life sciences, scores furnished by multiple observers on one or more targets, experiments and so on, are often used in various research. These ratings or scores are often subject to measurement error. Many research designs in studies of observer reliability give rise to categorical data via nominal scales (e.g., states of mental health such as normal, neurosis, and depression) or ordinal scales (e.g., stages of disease such as mild, moderate, and severe). In the normal situations, each of the observers classifies each subject once into exactly one category, taken from a fixed set of I categories. Many authors have discussed on these measures among which are Goodman and Kruskal (1954), Kendall and Stuart (1961, 1979), Somer(1962), Cohen (1960), Fleiss et al. (1969), Landis and Koch (1977a,b), Davies and Fleish (1982), Banerjee et al. (1999), Tanner and Young (1985a,b), Aickin (1990), Uebersax and Grove (1990), Agresti (1988, 1992), Agresti and lang (1993), Williamson and Manatunga (1997) and Barnhart and Williamson (2002), just but to mention a few.

In this paper we present some of the development so far achieved on these two measures as given by different authors and also to show with the aid of empirical example that agreement is a subset of association.

In section two and three we have the measures of association and agreement respectively. And in section four we present empirical examples on some of these measures that can handled $I > 2$ categories with general discussion of possible conclusion.

2 Measures of association

Association measures reflects the strength of the predictable relationship between the ratings of the two observers or raters. There are many indices that characterize the association between the row and column classifications of any I×I contingency table. If two observers or raters separately classify n subjects on I point scale, the resulting data can be summarized in the I×I table of observed proportions shown below: In this case $\pi_{ii'}$ is the proportion of subjects

Table 2.1: I×I table of observed proportion.

Obs1/Obs2	1	2	...	I	total
1	π_{11}	π_{12}	...	π_{1I}	π_{1+}
2	π_{21}	π_{22}	...	π_{2I}	π_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	...	π_{II}	π_{I+}
total	π_{+1}	π_{+2}	...	π_{+I}	1

classified into category i by observer 1 and into category i' by observer 2.

2.1 Basic measures of association

2.1.1 P coefficient

Kendall and Stuart (1961) proposed a coefficient of contingency due to Pearson denoted by

$$P = \{\frac{\chi^2}{n+\chi^2}\}^{\frac{1}{2}} \tag{2.1}$$

where χ^2 is the Pearson chi-square statistics for independence. This P coefficient ranges from 0 (for complete independence) to an upper limit of $(\frac{I-1}{I})^{\frac{1}{2}}$ (for perfect agreement) between the two observers. So the upper limit of this coefficient depends on the number of categories in the measurement scale.

2.1.2 T coefficient

In order to avoid this undesirable scale-dependency property of P above Tschuprow proposed an alternative function of χ^2 for I×I table, which is given in Kendall and Stuart (1961) as

$$T = \{\frac{\chi^2}{n(I-1)}\}^{\frac{1}{2}} \tag{2.2}$$

T ranges from 0 (for complete independence) to +1 (for perfect agreement) between the two observers. In the situation in which each of the two observers makes separate dichotomous judgements on n subjects, the resulting data can be summarized in the 2×2 table of observed proportions below: Thus, the

Table 2.2: 2×2 table of observed proportion.

Obs1/Obs2	0	1	total
0	π_{11}	π_{12}	π_{1+}
1	π_{21}	π_{22}	π_{2+}
total	π_{+1}	π_{+2}	1

relationship between the classifications of two observers can be characterized in terms of a contingency table measure of association.

2.1.3 Yule's Q coefficient

Kendall and Stuart (1961) also proposed a well known measure of association introduced by Yule (1900, 1912) in honor of Bulgarian statistician Quetelet, named *Yule's Q* denoted by

$$Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}}, \tag{2.3}$$

which ranges between -1 and +1 and has the following properties:

$$Q = \begin{cases} +1 & \text{if} \quad \pi_{12}\pi_{21} = 0 \quad i.e\ \pi_{12}\ or\ \pi_{21} = 0\ (and\ \pi_{11}, \pi_{22} > 0) \\ 0 & \text{if} \quad \pi_{11}\pi_{22} = \pi_{12}\pi_{21}, \quad i.e\ observers\ are\ independent \\ -1 & \text{if} \quad \pi_{11}\pi_{22} = 0, \quad i.e\ \pi_{11}\ or\ \pi_{22} = 0\ (and\ \pi_{12}, \pi_{21} > 0) \end{cases} \tag{2.4}$$

2.1.4 ϕ coefficient

Kendall and Stuart also proposed the ϕ coefficient denoted by

$$\phi = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2}} = \{\frac{\chi^2}{n}\}^{\frac{1}{2}}, \tag{2.5}$$

where χ^2 is the usual Pearson chi-square statistics for a 2×2 table. The ϕ coefficient ranges between -1 and +1 and the following properties:

$$\phi = \begin{cases} +1 & \text{if} \quad \pi_{12} = \pi_{21}, (and\ \pi_{12} < \pi_{11}, \pi_{22}) \quad i.e\ perfect\ agreement \\ 0 & \text{if} \quad \pi_{11}\pi_{22} = \pi_{12}\pi_{21}, \quad i.e\ observers\ are\ independent \\ -1 & \text{if} \quad \pi_{11} = \pi_{22}, (and\ \pi_{11} < \pi_{12}, \pi_{21}) \quad i.e\ complete\ disagreement \end{cases} \tag{2.6}$$

Both Q and ϕ measure the strength of the association between the classifications by the two observers. ϕ is not only a measure of association, but also a measure of agreement, since it reflects the extent to which the data cluster on the main diagonal of the table.

2.1.5 Gamma statistic

Another measure of association is the *gamma statistic* which was proposed by Goodman and Kruskal (1954). Given that the pair is untied on both variables, $\frac{\pi_c}{\pi_c+\pi_d}$ is the probability of concordance and $\frac{\pi_d}{\pi_c+\pi_d}$ is the probability of discordance. The difference between these probabilities is

$$\gamma = \frac{\pi_c - \pi_d}{\pi_c + \pi_d} \tag{2.7}$$

which is called gamma. The γ coefficient ranges between -1 and +1 and has the following properties:

$$\gamma = \begin{cases} +1 & \text{if} \quad \pi_d = 0 \\ 0 & \text{if} \quad \pi_c = \Pi_d \\ -1 & \text{if} \quad \pi_c = 0 \end{cases} \tag{2.8}$$

The probability of concordance and discordance π_c and π_d, that is, the probability that a pair of observations is concordant or discordant respectively, can be shown to be

$$\pi_c = 2\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\pi_{ij}\left(\sum_{k>i}\sum_{l>j}\pi_{kl}\right) \tag{2.9}$$

$$\pi_d = 2\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\pi_{ij}\left(\sum_{k>i}\sum_{l\leq j}\pi_{kl}\right) \tag{2.10}$$

The factor 2 occurs in these formulas because the first observation could be in cell (i, j) and the second in cell (k, l), or vice versa. We can see that π_c and π_d are sums of products of sums of probabilities, and can be written using the $\exp - \log$ notation in the Appendix (Forthofer and Koch, 1973; and Bergsma, 1997).

Also the probabilities of a tie on the variables involve say, A, B, and both A and B are

$$\pi_{t,A} = \sum_i(\pi_{i+})^2 \quad \pi_{t,B} = \sum_j(\pi_{+j})^2 \quad \pi_{t,AB} = \sum_{ij}\pi_{ij}^2 \tag{2.11}$$

2.1.6 Somer's-d statistic

Somer(1962) also proposed another statistic for measuring association which is similar to gamma, but for which the pairs untied on one variable $(1 - \pi_{t,A})$ or $(1 - \pi_{t,B})$ rather than on both variables $(\pi_c + \pi_d)$. The population value of the statistic is given as

$$\Delta_{BA} = \frac{\pi_c - \pi_d}{1 - \pi_{t,A}} \tag{2.12}$$

This expression is the difference between the proportions of concordant and discordance pairs out of the pairs that are untied on A. This is an asymmetry measure intended for use when B is a response variable.

2.1.7 Kendall's tau-b and tau

Kendall (1945) proposed another statistic called Kendall's tau-b which is given as

$$\tau_b = \frac{\pi_c - \pi_d}{\sqrt{(1 - \pi_{t,A})(1 - \pi_{t,B})}} \tag{2.13}$$

If there are no ties, the common value of gamma, Somers'd, and Kendall's tau-b is

$$\tau = \pi_c - \pi_d \tag{2.14}$$

This measure is refers to as Kendall's tau and originally introduced for continuous variables.

2.1.8 Association coefficient for nominal scales

Kendall and Stuart (1979) proposed another measure of association mainly for nominal scales. let V(Y) denote a measure of variation for the marginal distribution $\{\pi_{+1}, ..., \pi_{+I}\}$ of the response Y, and let $V(Y|i)$ denote this measure computed for the conditional distribution $\{\pi_{1|i}, ..., \pi_{I|i}\}$ of Y at the i^{th} setting of an explanatory variable X. A proportional reduction in variation measure, has form

$$\frac{V(Y) - E[V(Y|X)]}{V(Y)} \tag{2.15}$$

where $E[V(Y|X)]$ is the expectation of the conditional variation taken with respect to the distribution of X. When X is a categorical variable having marginal distribution $\{\pi_{1+}, ..., \pi_{I+}\}$, $E[V(Y|X)] = \sum_i \pi_{i+} V(Y|i)$.

2.1.9 Concentration coefficient

Goodman and Kruskal (1954) proposed another coefficient for measuring association in a contingency table called τ, which can be used for tables on nominal scales based on what described in section (2.1.8). Let

$$V(Y) = \sum_j \pi_{+j}(1 - \pi_{+j}) = 1 - \sum \pi_{+j}^2.$$

This gives the probability that two independent observations from the marginal distribution of Y falls in different categories. The conditional variation in row i is then

$$V(Y|i) = 1 - \sum \pi_{j|i}^2.$$

The average conditional variation for an I×J table with joint probabilities $\{\pi_{ij}\}$ is

$$E[V(Y|X)] = 1 - \sum_i \pi_{i+} \sum_j \pi_{j|i}^2 = 1 - \sum \pi_{ij}^2 / \pi_{i+}.$$

Therefore, the proportional reduction in variation is Goodman and Kruskal's tau

$$\tau = \frac{\sum_{ij} \pi_{ij}^2/\pi_{i+} - \sum \pi_{+j}^2}{1 - \sum \pi_{+j}^2} \tag{2.16}$$

which is also called the *concentration coefficient.* $0 \leq \tau \leq 1$.

2.1.10 Uncertainty coefficient

Another alternative measure to (2.16) called *uncertainty coefficient* was also proposed by Theil (1970), which is denoted as U below

$$U = \frac{\sum_{ij} \pi_{ij} \log(\pi_{ij}/(\pi_{i+}\pi_{+j}))}{\sum_j \pi_{+j} \log(\pi_{+j})} \tag{2.17}$$

The measures of τ and U are well defined when more than one $\pi_{+j} > 0$. Also $0 \leq U \leq 1$. $\tau = U = 0$ implies independence of the two variables X and Y and $\tau = U = 1$ means no conditional variation.

2.1.11 Pearson's correlation coefficient

A useful measure of association for two interval level variables when these variables are linearly related is Pearson's correlation coefficient denoted by ρ, which is defined as

$$\rho = \frac{cov(A,B)}{\sigma_A \sigma_B} = \frac{E(AB) - E(A)E(B)}{\sigma_A \sigma_B} \tag{2.18}$$

Let π_{ij} be the cell probability for cell (i,j). The $E(A) = \sum_i a_i \pi_{i+}$ and $E(B) = \sum_j b_j \pi_{+j}$, where a_i and b_j are scores of categories I of A and J of B respectively. Also let $E(AB) = \sum_i \sum_j a_i b_j \pi_{ij}$. Therefore, ρ is a sum of products of sums of products of sums of probabilities.

2.1.12 Odds Ratio

Another measure of association is called the *Odds Ratio.* Given a 2×2 contingency table of the form in table 2 above, the probability of success is π_{11} in row 1 and π_{21} in row 2. Within row 1 and row 2, the odds of successes denoted by α are defined to be

$$\alpha_1 = \frac{\pi_{11}}{\pi_{1+} - \pi_{11}} = \frac{\pi_{11}}{\pi_{12}} \tag{2.19}$$

$$\alpha_2 = \frac{\pi_{21}}{\pi_{2+} - \pi_{21}} = \frac{\pi_{21}}{\pi_{22}} \tag{2.20}$$

respectively. The odds α are nonnegative with value greater than 1.0 when a success is more likely than a failure. When odds $\alpha = 4.0$, a success is four times as likely as a failure. In either row, the success probability is the function of the odds, this can be obtained by making π (probability of success) the subject of formula in each of the above equations, that is,

$$\pi = \frac{\alpha}{\alpha + 1} \tag{2.21}$$

The ratio of odds from the two rows,

$$\theta = \frac{\alpha_1}{\alpha_2} = \frac{\frac{\pi_{11}}{\pi_{12}}}{\frac{\pi_{21}}{\pi_{22}}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{2.22}$$

is called the *Odds ratio.* The θ coefficient ranges between 0 and ∞ and has the following properties:

$$\theta = \begin{cases} +1 & \text{if} \quad \pi_{11}\pi_{22} = \pi_{12}\pi_{21} \quad i.e\ \alpha_1 = \alpha_2 \\ > 1\ and\ < \infty & \text{if} \quad \pi_{11}\pi_{22} > \pi_{12}\pi_{21} \quad i.e\ \alpha_1 > \alpha_2 \\ > 0\ and\ < 1 & \text{if} \quad \pi_{11}\pi_{22} < \pi_{12}\pi_{21} \quad i.e\ \alpha_1 < \alpha_2 \end{cases} \tag{2.23}$$

Values of θ farther from 1.0 in given direction represent stronger levels of association.

3 Measures of Agreement

Agreement is a special case of association which reflects the extent to which observers classify a given subject identically into the same category. In order to assess the psychometric integrity of different ratings we compute interrraters reliability and/ or interrater agreement.

Interrater reliability coefficients reveal the similarity or consistency of the pattern of responses, or the rank-ordering of responses between two or more raters (or two or more rating sources), independent of the level or magnitude of those ratings. For example, let us consider the following table, One can observe from

Table 3.1: Ratings of three subjects by three raters.

subject	Rater 1	Rater 2	Rater 3
1	5	6	2
2	3	4	2
3	1	2	1

the table that all the raters were consistent in their ratings, rater 2 maintained his leading ratings followed by rater 1 and rater 3 respectively.
Interrater agreement on the other hand is to measure the degree that ratings are similar in level or magnitude. It pertains to the extent to which the raters classify a given subject identically into the same category. Kozlowski and Hattrup (1992) noted that an interrater agreement index is designed to "reference the interchangeability among raters; it addresses the extent to which raters make essentially the same ratings". Thus, theoretically, obtaining high levels of agreement should be more difficult than obtaining high levels of reliability or consistency. Also consider the table below, From Table 3.2 one can observe

Table 3.2: Ratings of three subjects by three raters.

subject	Rater 1	Rater 2	Rater 3
1	5	5	3
2	3	3	2
3	1	1	1

that the ratings are similar compare to Table 3.1.

3.1 Basic measures of Agreement

3.1.1 *Cohen's Kappa coefficient:*

Cohen (1960) proposed a standardized coefficient of raw agreement for nominal scales in terms of the proportion of the subjects classified into the same category by the two observers, which is estimated as

$$\pi_o = \sum_{i=1}^{I} \pi_{ii} \tag{3.1}$$

and under the baseline constraints of complete independence between ratings by the two observers,which is the expected agreement proportion estimated as

$$\pi_e = \sum_{i=1}^{I} \pi_{i.}\pi_{.i} \tag{3.2}$$

The Kappa statistic can now be estimated by

$$\widehat{k_c} = \frac{\widehat{\pi}_o - \widehat{\pi}_e}{1 - \widehat{\pi}_e} \tag{3.3}$$

where $\widehat{\pi}_o$ and $\widehat{\pi}_e$ are as defined above.
Early approaches to this problem have focused on the observed proportion of agreement (Goodman and Kruskal 1954), thus suggesting that chance agreement can be ignored. Later Cohen's kappa was introduced for measuring nominal scale chance-corrected agreement. Scott (1955) defined π_e using the underlying assumption that the distribution of proportions over the I categories for the population is known, and is equal for the two raters. Therefore if the two raters are interchangeable, in the sense that the marginal distributions are identical, then Cohen's and Scott's measures are equivalent because Cohen's kappa is an extension of Scott's index of chance-corrected measure. To determine whether $\widehat{k}$ differs significantly from zero, one could use the asymptotic variance formulae given by Fleiss et al. (1969) for the general I×I tables. For large n, Fleiss et al.'s formulae is practically equivalent to the exact variance derived by Everitt (1968) based on the central hypergeometric distribution. Under the hypothesis of only chance agreement, the estimated large-sample variance of $\widehat{k}$ is given by

$$\widehat{var}_o(\widehat{k_c}) = \frac{\pi_e + \pi_e^2 - \sum_{i=1}^{I} \pi_{i.}\pi_{.i}(\pi_{i.} + \pi_{.i})}{n(1-\pi_e)^2}. \tag{3.4}$$

Assuming that

$$\frac{\widehat{k}}{\sqrt{\widehat{var}_o(\widehat{k})}} \tag{3.5}$$

follows a normal distribution, one can test the hypothesis of chance agreement by reference to the standard normal distribution. In the context of reliability studies, however, this test of hypothesis is of little interest, since generally the raters are trained to be reliable. In this case, a lower bound on kappa is more appropriate. This requires estimating the nonnull variance of $\widehat{k}$, for which Fleiss et al. provided an approximate asymptotic expression, given by:

$$\begin{aligned}\widehat{var}(\widehat{k}) &= \frac{1}{n(1-\pi_e)^2}\left(\sum_{i=1}^{I}\pi_{ii}\{1-(\pi_{i.}+\pi_{.i})(1-\widehat{k})\}^2+(1-\widehat{k})^2\right)\\ &\times\left(\sum_{i\neq i'}^{I}\pi_{ii'}(\pi_{i.}+\pi_{.i'})^2-\{\widehat{k}-\pi_e(1-\widehat{k})\}^2\right). \qquad (3.6)\end{aligned}$$

Fleiss (1971) proposed a generalization of Cohen's kappa statistic to the measurement of agreement among a constant number of raters (say, K). Each of the n subjects are related by K (>2) raters independently into one of m mutually exclusive and exhaustive nominal categories. This formulation applies to the case of different sets of raters (that is random ratings) for each subject. The motivated example is a study in which each of 30 patients was rated by 6 psychiatrists (selected randomly from a total pool of 43 psychiatrists) into one of five categories.

Let k_{ij} be the number of raters who assigned the ith subject to the jth category $i = 1, 2, ..., n$, $j = 1, 2, ..., m$ and define

$$\pi_j = \frac{1}{Kn}\sum_{i=1}^{n} K_{ij} \qquad (3.7)$$

π_j is the proportion of all assignments which were to the jth category. The chance corrected measure of overall agreement proposed by Fleiss (1971) is given by

$$\widehat{k} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}K_{ij}^2 - Kn\{1+(K-1)\sum_{j=1}^{m}\pi_j^2\}}{nK(K-1)(1-\sum_{j=1}^{m}\pi_j^2)} \qquad (3.8)$$

Under the null hypothesis of no agreement beyond chance, the K assignments on one subject are multinomial variables with probabilities $\pi_1, \pi_2, ..., \pi_m$. Using this Fleiss (1971) obtained an approximate asymptotic variance of $\widehat{k}$ under the hypothesis of no agreement beyond chance:

$$var_o\widehat{k} = A\left\{\frac{\sum_{j=1}^{m}\pi_j^2-(2K-3)(\sum_{j=1}^{m}\pi_j^2)^2+2(K-2)\sum_{j=1}^{m}\pi_j^3}{(1-\sum_{j=1}^{m}\pi_j^2)^2}\right\} \qquad (3.9)$$

where

$$A = \frac{2}{nK(K-1)}.$$

Apart from $\widehat{k}$ statistic for measuring overall agreement, Fleiss (1971) also proposed a statistic to measure the extent of agreement in assigning a subject to a particular category. A measure of the beyond chance agreement in assignment to category given by

$$\widehat{k_j} = \frac{\sum_{i=1}^{n}K_{ij}^2 - Kn\pi_j\{1+(K-1)\pi_j\}}{nK(K-1)\pi_j(1-\pi_j)} \qquad (3.10)$$

The measure of overall agreement $\widehat{k}$ is a weighted average of $\widehat{k_j}$'s, with the corresponding weights $\pi_j(1-\pi_j)$. The approximate asymptotic variance of $\widehat{k_j}$ under the null hypothesis of no agreement beyond chance is

$$var_o\widehat{k_j} = \frac{\{1+2(K-1)\pi_j\}^2 + 2(K-1)\pi_j(1-\pi_j)}{nK(K-1)^2\pi_j(1-\pi_j)} \tag{3.11}$$

Landis and Koch (1977a) have characterized different ranges of arbitrary values for kappa with respect to the degree of agreement they suggest and these have become a standard in all the literatures,see below the the ranges of kappa statistic with the respective strength of agreement:
There is a wide disagreement about the usefulness of kappa statistic to assess

Table 3.3: Range of kappa statistic with the respective strength of agreement.

Kappa statistic	*Strength of agreement*
< 0.00	poor
0.00-0.20	slight
0.21-0.40	fair
0.41-0.60	moderate
0.61-0.80	substantial
0.81-1.00	almost perfect.

rater agreement (Maclure and Willett 1987 and 1988). At least, it can be said that

- kappa statistic should not be viewed as the unequivocal standard or default way to quantify agreement;
- one should be concerned about using a statistic that is the source of so much controversy;
- one should consider alternatives and make an informed choice.

One can distinguish between two possible uses of kappa (Thompson and Walter 1988a and 1988b, Kraemer and Bloch 1988, Guggenmoos-Holzmann 1993),

(i) as a way to test rater independence, that is, as a test statistics, which involves testing the null hypothesis that there is no more agreement than might occur by chance given random guessing; that is, one makes a qualitative, "yes or no" decision about whether raters are independent or not. Kappa is appropriate for this purpose, although to know that raters are not independent is not very informative; raters are dependent by definition, inasmuch as they are rating the same cases.

(ii) as a way to quantify the level of agreement, that is, as an effect-size measure, which is the source of concern. Kappa's calculation uses a term called the proportion of chance (or expected) agreement. This is interpreted as the proportion of times raters would agree by chance alone. However, the term is relevant only under the conditions of statistical independent of raters. Since raters are clearly not independent, the relevance of this term, and its appropriateness as a correction to actual agreement levels, is very questionable.

Thus, the common statement that kappa is a "chance-corrected measure of agreement" (Landis and Koch 1977b; Davies and Fleish 1982; Banerjee et al. 1999) is misleading. As a test statistic, kappa can verify that agreement exceeds chance levels. But as a measure of the level of agreement, kappa is not "chance-corrected"; indeed, in the absence of some explicit model of rater decision making, it is by no means clear how chance affects the decisions of actual raters and how one might correct for it. A better case for using kappa to qualify rater agreement is that, under certain conditions, it approximates the intra-class correlation. But this too is problematic in that (1) these conditions are not always met, and (2) one could instead directly calculate the intra-class correlation.

3.1.2 *Weighted Kappa coefficient:*

Cohen (1968) proposed a modified form of kappa called Weighted kappa which allows for scales disagreement or partial credit. Often situations arise when certain disagreements between two raters are more serious than others. For example, in an agreement study of psychiatric diagnosis in the categories personality disorder, neurosis and psychosis, a clinician would likely consider a diagnostic disagreement between neurosis and psychosis to be more serious than between neurosis and personality disorder. However, $\widehat{k}$ makes no such distinction, implicitly treating all disagreements equally. Weighted Kappa is defined as

$$\widehat{k_w} = \frac{\pi_o^* - \pi_e^*}{1 - \pi_e^*} \tag{3.12}$$

where

$$\pi_o^* = \sum_{i=1}^{I} \sum_{i'=1}^{I} w_{ii'} \pi_{ii'} \tag{3.13}$$

and

$$\pi_e^* = \sum_{i=1}^{I} \sum_{i'=1}^{I} w_{ii'} \pi_{i.} \pi_{.i'} \tag{3.14}$$

where $\{w_{ii'}\}$ is the weights, which in most cases $0 \leq w_{ii'} \leq 1$ for all i, i', so that π_o^* is a weighted observed proportion of agreement, and π_e^* is the corresponding weighted proportion of agreement expected under the constraints of total independence. Note that the Unweighted kappa is a special case of $\widehat{k_w}$ with $w_{ii'} = 1$ for $i = i'$ and $w_{ii'} = 0$ for $i \neq i'$. Also if the I categories form an ordinal scale, with the categories assigned the numerical values $1, 2, ..., I$, and

$$w_{ii'} = 1 - \frac{(i - i')^2}{(I - 1)^2}, \tag{3.15}$$

then $\widehat{k_w}$ can be interpreted as an intra-class correlation coefficient for a two-way ANOVA computed under the assumption that the n subjects and the two raters are random samples from populations of subjects and raters, respectively (Fleiss and Cohen, 1973).

Fleiss et al.(1969) calculated the unconditional large sample variance of weighted kappa as

$$\begin{aligned}\widehat{var}(\widehat{k_w}) &= \frac{1}{n(1-\pi_e^*)^4}(\sum_{i=1}^{I}\sum_{i'=1}^{I}\pi_{ii'}[w_{ii'}(1-\pi_e^*) \\ &\quad -(\overline{w}_{i.} + \overline{w}_{.i'})(1-\pi_o^*)]^2 \\ &\quad -(\pi_o^*\pi_e^* - 2\pi_e^* + \pi_o^*)^2)\end{aligned} \tag{3.16}$$

where

$$\overline{w}_{i.} = \sum_{i'=1}^{I} w_{ii'}\pi_{.i'} \ and \ \overline{w}_{.i'} = \sum_{i=1}^{I} w_{ii'}\pi_{i.}. \tag{3.17}$$

Cicchetti (1972) recommended another weights as

$$w_{ii'} = 1 - \frac{| \, i - i' \, |}{(I - 1)}, \tag{3.18}$$

Cicchetti used these weights to test for the significance of observer agreement through the Cicchetti test statistic Z_c

$$Z_c = \frac{\pi_o^* - \pi_e^*}{[\widehat{var(\pi_o^*)}]^{\frac{1}{2}}} \tag{3.19}$$

where

$$[\widehat{var(\pi_o^*)}] = \frac{1}{n-1}\left[\sum_{i=1}^{I}\sum_{i'=1}^{I} w_{ii'}^2\pi_{ii'} - \pi_o^{*2}\right] \tag{3.20}$$

Cohen (1968) has shown that under observed marginal symmetry, weighted kappa $\widehat{k_w}$ is precisely equal to the product-moment correlation by choosing the weights to be

$$w_{ii'} = 1 - (i - i')^2, \tag{3.21}$$

when the I categories are not only ordinal scale, but also assumed equal spaced along some underlying continuum. Discrete numerical integers such as $1, 2, ..., I$ can then be assigned to the respective classes (Barnhart and Williamson 2002).

Oden (1991) proposed a method to estimate a pooled kappa between two raters when both raters rate the same set of pairs of the body like eyes. His method assumes that the true left-eye and right-eye kappa values are equal and makes use of the correlated data to estimate confidence intervals for the common kappa.

The pooled kappa estimator is a weighted average of the kappas for the right and left eyes. We define letters B and D as follow

$$B = (1 - \sum_{1=1}^{m}\sum_{1=1}^{m} w_{ij}\rho_{i.}\rho_{.j})\widehat{k}_{right} + (1 - \sum_{1=1}^{m}\sum_{1=1}^{m} w_{ij}\lambda_{i.}\lambda_{.j})\widehat{k}_{left}$$

$$D = (1 - \sum_{1=1}^{m}\sum_{1=1}^{m} w_{ij}\rho_{i.}\rho_{.j}) + (1 - \sum_{1=1}^{m}\sum_{1=1}^{m} w_{ij}\lambda_{i.}\lambda_{.j})$$

so that the pool kappa will be the ratio of the two letters,

$$\widehat{k}_{pooled} = \frac{B}{D} \tag{3.22}$$

where ρ_{ij}=proportion of patients whose right eye was rated i by rater 1 and j by rater 2,
λ_{ij}=proportion of patients whose left eye was rated i by rater 1 and j by rater 2,
w_{ij}=agreement weight that reflects the degree of agreement between raters 1 and 2 if they use rating i and j respectively for the same eye,
and

$$\rho_{i.},\ \rho_{.j},\ \lambda_{i.},\ \lambda_{.j}$$

have their usual meanings. By applying the delta method, Oden obtained an approximate standard error of the pool kappa estimator.

Schouten (1993) also proposed another alternative method for paired data situation. He noted that the Cohen (1968); Fleiss et al. (1969) weighted kappa formula and its standard error can be used if the observed as well as the chance agreement is averaged over the two sets of eyes and then substituted into the formula for kappa. To this end, let each eye be diagnosed normal or abnormal, and let each patient be categorized into one of the following four categories by each rater:

R+L+: abnormality is present in both eyes

R+L-: Abnormality is present in the right eye but not in the left eye

R-L+: Abnormality is present in the left eye but not in the right eye

R-L-: Abnormality is absent in both eyes

The frequencies of the ratings can be presented as follows: Schouten (1993)

used the weighted kappa statistic to determine an overall agreement measure.

Table 3.4: Binocular data frequencies and agreement weights.

Category	rater 2				
rater 1	R+L+	R+L-	R-L+	R-L-	Total
R+L+	$n_{11}(1.0)$	$n_{12}(0.5)$	$n_{13}(0.5)$	$n_{14}(0.0)$	$n_{1.}$
R+L-	$n_{21}(0.5)$	$n_{22}(1.0)$	$n_{23}(0.0)$	$n_{24}(0.5)$	$n_{2.}$
R-L+	$n_{31}(0.5)$	$n_{32}(0.0)$	$n_{33}(1.0)$	$n_{43}(0.5)$	$n_{3.}$
R-L-	$n_{41}(0.0)$	$n_{42}(0.5)$	$n_{43}(0.5)$	$n_{44}(1.0)$	$n_{4.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

He defined the agreement weights w_{ij} which are represented in parenthesis in the table above as

$$w_{ij} = \begin{cases} 1.0 & \text{for} \quad Complete\ agreement,\ i.e\ if\ the\ raters\ agreed\ on\ both\ eyes \\ 0.5 & \text{for} \quad partial\ agreement, i.e\ if\ one\ agreed\ and\ one\ disagreed \\ 0.0 & \text{for} \quad Complete\ disagreement,\ i.e\ if\ the\ raters\ disagreed\ on\ both\ eyes \end{cases} \tag{3.23}$$

The overall agreement measure is then defined to be

$$\widehat{k_w} = \frac{\pi_o^{**} - \pi_e^{**}}{1 - \pi_e^{**}} \tag{3.24}$$

where

$$\pi_o^{**} = \frac{\sum_{i=1}^{4} \sum_{j=1}^{4} w_{ij} n_{ij}}{n} \tag{3.25}$$

and

$$\pi_e^{**} = \frac{\sum_{i=1}^{4} \sum_{j=1}^{4} w_{ij} n_{i.} n_{.j}}{n^2} \tag{3.26}$$

The standard error can be calculated as Fleiss et al. (1969). This can be extended for more than two raters by simply adjusting the agreement weights. Shoukri et al. (1995) proposed another method of agreement measure when the pairing situation is such that raters classify individuals blindly by two different rating protocols into one of two categories, such as to establish the congruent validity of the two rating protocols. For example, as stated by Banerjee et al. (1999), consider two tests for routine diagnosis of paratuberculosis in cattle animals which are the dot immunobinding assay (DIA) and the enzyme linked immunosorbent assay (ELISA). Comparison of the results of these two tests depends on the serum samples obtained from the cattle. One can then evaluate the same serum sample using both tests, a procedure that clearly creates a realistic "matching". Let

$$X_i = \begin{cases} 1.0 & \text{if} \quad ith\ serum\ sample\ tested\ by\ DIA\ is\ positve. \\ 0.0 & \text{if} \quad ith\ serum\ sample\ tested\ by\ DIA\ is\ negative. \end{cases} \tag{3.27}$$

and let

$$Y_i = \begin{cases} 1.0 & \text{if} \quad ith\ serum\ sample\ tested\ by\ ELISA\ is\ positve. \\ 0.0 & \text{if} \quad ith\ serum\ sample\ tested\ by\ ELISA\ is\ negative. \end{cases} \tag{3.28}$$

Let π_{kl} $(k, l = 0, 1)$ denote the probability that $X_i = k$ and $Y_i = l$. Then $\pi_1 = \pi_{11} + \pi_{01}$ is the probability that a serum sample tested by ELISA is positive, and $\pi_2 = \pi_{11} + \pi_{10}$ is the probability that the matched serum sample tested by DIA is positive. Under this model, kappa reduces to the following expression:

$$k = \frac{2\rho\{\pi_1(1-\pi_1)\pi_2(1-\pi_2)\}^{\frac{1}{2}}}{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)} \tag{3.29}$$

where ρ is the correlation coefficient between X and Y. As a result of this random sample of n pairs of correlated binary responses, Shoukri et al. obtained the maximum likelihood estimate of k as

$$\widehat{k} = \frac{2(\bar{t}-\bar{x}\bar{y})}{\bar{y}(1-\bar{x})+\bar{x}(1-\bar{y})} \tag{3.30}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \bar{t} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$; the asymptotic variance was also obtained for this corrected to the first order of approximation. Using the large sample variance expression, one could test the hypothesis that the two diagnostic tests are uncorrelated.

For more on weighted and unweighted Kappa see Barlow et al.(1991); Schouten (1993); Shoukri et al. (1995); Donner et al. (1996); Banerjee et al. (1999), Gonin et al. (2000), and Barnhart and Williamson (2002), Shoukri (2004).

3.1.3 *Intraclass kappa*

Intraclass kappa was defined for data consisting of blinded dichotomous ratings on each of n subjects by two fixed raters. It is assumed that the ratings on a subject are interchangeable; that is in the population of subjects, the two ratings for each subject have a distribution that is invariant under permutations of the raters to ensure that there is no rater bias (Scott (1955), Bloch and Kraemer (1989), Donner and Eliasziw (1992),Banerjee et al. (1999), Barnhart and Williamson (2002). Let X_{ij} denote the rating for the ith subject by the jth rater, $i = 1, 2, ..., n$, $j = 1, 2$. and for each subject i, let $\pi_i = P(X_{ij=1})$ be the probability that the rating is a success. Over the population of subjects, let $E(\pi_i) = \Pi$, $\Pi' = 1 - \Pi$ and $var(\pi_i) = \sigma_\pi^2$. The intraclass kappa as defined by Bloch and Kraemer (1989) is then

$$k_I = \frac{\sigma_\pi^2}{\Pi\Pi'} \tag{3.31}$$

To obtain the estimator of intraclass kappa, let us consider the the following table of probability model for joint responses with kappa coefficient explicitly defined in its parametric structure.
Thus, the log-likelihood function is given by

$$\begin{aligned}\log L(\Pi, k_I \setminus n_{11}, n_{12}, n_{21}, n_{22}) \quad &= \quad n_{11}\log(\pi^2 + k_I\Pi\Pi') + (n_{12}+n_{21}) \\ &\quad\quad \log\{\Pi\Pi'(1-k_I)\} + n_{22}\log(\pi'^2 + k_I\Pi\Pi').\end{aligned}$$

Table 3.5: Underlying model for estimation of intraclass kappa

X_{i1}	X_{i2}	Observed frequency	Expected Probability.
1	1	n_{11}	$\pi^2 + k_I \Pi\Pi'$
1	0	n_{12}	$\Pi\Pi'(1-k_I)$
0	1	n_{21}	$\Pi\Pi'(1-k_I)$
0	0	n_{22}	$\pi'^2 + k_I \Pi\Pi'$

The maximum likelihood estimators $\widehat{\pi}$ and $\widehat{k}_I$ for Π and k_I are obtained as

$$\widehat{\pi} = \frac{2n_{11} + n_{12} + n_{21}}{2n}, \tag{3.32}$$

and

$$\widehat{k}_I = \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})}, \tag{3.33}$$

with the estimated standard error for $\widehat{k}_I$ given by Block and Kraemer (1989)

$$SE(\widehat{k}_I) = \{\frac{1-\widehat{k}_I}{n}[(1-\widehat{k}_I)(1-2\widehat{k}_I) + \frac{\widehat{k}_I(2-\widehat{k}_I)}{2\widehat{\pi}(1-\widehat{\pi})}]\}^{\frac{1}{2}}. \tag{3.34}$$

and with this $100(1-\alpha)\% = \widehat{k}_I \pm Z_{1-\frac{\alpha}{2}} SE(\widehat{k}_I)$ confidence interval can be obtained for $\widehat{k}_I$. This has reasonable properties only in a very large samples that are not typical of of the size of the most interrater agreement studies.

Barnhart and Williamson (2002) considered intraclass kappa for measuring agreement between two readings for a categorical response with I categories if the two readings are replicated measurements. It assumes no bias because the probability of a positive rating is the same for the two readings due to replication, and it is given as

$$k_{In} = \frac{\sum_{i=1}^{I} \pi_{ii} - \sum_{i=1}^{I}((\pi_{i+} + \pi_{+i})/2)^2}{1 - \sum_{i=1}^{I}((\pi_{i+} + \pi_{+i})/2)^2} \tag{3.35}$$

Donner and Eliasziw (1992) proposed a procedure based on chi-square goodness of fit statistic to construct confidence interval for small samples. This was done by equating the computed one degree of freedom chi-square statistic to an appropriately selected critical value, and solving for the two roots of kappa. The upper $\widehat{k}_U$ and the lower $\widehat{k}_L$ limits of $100(1-\alpha)\%$ confidence interval for $\widehat{k}_I$ are obtained as

$$\widehat{k_L} = (\frac{1}{9}y_3^2 - \frac{1}{3}y_2)^{\frac{1}{2}}(\cos\frac{\theta + 2\pi}{3} + \sqrt{3}\sin\frac{\theta + 2\pi}{3}) - \frac{1}{3}y_3 \tag{3.36}$$

$$\widehat{k_U} = 2(\frac{1}{9}y_3^2 - \frac{1}{3}y_2)^{\frac{1}{2}}\cos\frac{\theta + 5\pi}{3} - \frac{1}{3}y_3, \tag{3.37}$$

where $\pi = 3.14159$, $\theta = \arccos\frac{V}{W}$, $V = \frac{1}{27}y_3^3 - \frac{1}{6}(y_2y_3 - 3y_1)$, $W = (\frac{1}{9}y_3^2 - \frac{1}{3}y_2)^{\frac{1}{2}}$; and

$$y_1 = \frac{\{n_{12} + n_{21} - 2n\widehat{\Pi}(1-\widehat{\Pi})\}^2 + 4n^2\widehat{\Pi}^2(1-\widehat{\Pi})^2}{4n\widehat{\Pi}^2(1-\widehat{\Pi})^2(\chi^2_{1,1-\alpha} + n)} - 1, \tag{3.38}$$

$$y_2 = \frac{(n_{12}+n_{21})^2 - 4n\widehat{\Pi}(1-\widehat{\Pi})\{1-4\widehat{\Pi}(1-\widehat{\Pi})\}\chi^2_{1,1-\alpha}}{4n\widehat{\Pi}^2(1-\widehat{\Pi})^2(\chi^2_{1,1-\alpha}+n)} - 1, \tag{3.39}$$

$$y_3 = \frac{n_{12}+n_{21}+\{1-2\widehat{\Pi}(1-\widehat{\Pi})\}\chi^2_{1,1-\alpha}}{\widehat{\Pi}(1-\widehat{\Pi})(\chi^2_{1,1-\alpha}+n)} - 1. \tag{3.40}$$

Donner and Eliasziw (1992) also describe hypothesis-testing and sample size calculations using this goodness of fit procedure. Donner and Eliasziw (1997) has also extended the above to a case of three or more rating categories per subject.

Barlow (1996) extended the intraclass kappa to accommodate subject-specific covariates directly in the model. Although both raters have the same marginal probability of classification, this probability is assumed to be a function of the covariates. Barlow used a trinomial model obtained by collapsing the two discordance cells in table 7 into a single cell. The ratings of each subject are placed in one of three classification cells (both success,discordant, and both failure). Let Y_{ik} be an indicator of the placement of subject i in cell $k = 1, 2, 3$. For example, if for subject i both ratings were success, then $y_{i1} = 1$ and $y_{i2} = y_{i3} = 0$. Also let $X_i = (1, X_{i1}, X_{i2}, ..., X_{ip})$ be the vector of covariates for subject i. Assuming a logit link function between the mean π_i and the covariate vector X_i, that is, $\log\{\frac{\pi_i}{(1-\pi_i)}\} = X_i\beta$, where β is the parameter vector to be estimated. Then the multinomial likelihood is given by

$$\begin{aligned} L(\beta, k_l \mid X, Y) \quad \propto \quad & \prod_{i=1}^{n} \frac{e^{X_i\beta}}{(1+e^{X_i\beta})^2}\{e^{X_i\beta}+k_l\}^{y_{i1}}\{2(1-k_l)\}^{y_{i2}} \\ & \times\{e^{-X_i\beta}+k_l\}^{y_{i3}}. \end{aligned} \tag{3.41}$$

This function is hard to maximize; however, Barlow noted that it is equivalent to the likelihood of a conditional logistic regression model with a general relative risk function r and one case ($y_{ik} = 1$) and two controls ($y_{ij} = 0,\ j \neq k$) defined for each subject. Specifically, the relative risk r_i can be expressed as $r_i = e^{z_i\beta} + w_i k_l - (w_i - 1)/3$, where

$$z_i = \begin{cases} X_i & \text{If} \quad Y_{i1} = 1, \\ 0.0 & \text{if} \quad Y_{i2} = 1, \\ -X_i & \text{If} \quad Y_{i3} = 1. \end{cases} \tag{3.42}$$

and

$$w_i = \begin{cases} 1 & \text{If} \quad Y_{i1} = 1, \\ -2 & \text{if} \quad Y_{i2} = 1, \\ 1 & \text{If} \quad Y_{i3} = 1. \end{cases} \tag{3.43}$$

The additive risk function decomposes the risk into a part that incorporates the covariate as a part that depends on the intraclass kappa, and an "offset" that is 0 for concordant observations and 1 for disconcordant observations. The above model can be fitted using any suitable software. In addition, in getting estimates for k_l and β, standard errors and Wald confidence intervals are obtained.

In a situation where we have multiple trial of an experiment in different regions or centers, in each of the center a reliability studies has to be conducted and this will give rise to several independent kappa statistics which can be used to test for homogeneity across the centers or regions, that is testing $H_o : k_1 = k_2 = ...k_N$, where k_h denoted the population kappa value for center h. Donner et al (1996) proposed methods of testing homogeneity of N independent kappa of the intraclass form. Their underlying model assumed that N independent studies, involving $n = \sum_{h=1}^{N} n_h$ subjects, have been completed, where each subject is given a dichotomous rating (success-failure) by each of the two raters. In addition, it is assumed that the marginal probability (Π_h) of classifying a subject as success constant across raters in a particular study; but this probability may varies across the N studies, which means there is no rater bias within the studies. The probabilities of joint responses within study h arise from a trinomial model which can be obtained by collapsing the two discordant cells in table 7 into a single cell as follows:

$\pi_{1h}(k_h) = \Pi_h^2 + \Pi_h(1 - \Pi_h)k_h$, (both successes).

$\pi_{2h}(k_h) = 2\Pi_h(1 - \Pi_h)(1 - k_h)$, (one success and one failure),

$\pi_{3h}(k_h) = (1 - \Pi_h)^2 + \Pi_h(1 - \Pi_h)k_h$, (both failure).

These are the same expression as presented in table 7, with the exception of Π_h being study specific. For the hth study, maximum likelihood estimators for Π_h and k_h are given by

$$\widehat{\Pi}_h = \frac{2n_{1h} + n_{2h}}{2n_h}, \tag{3.44}$$

and

$$\widehat{k}_h = 1 - \frac{n_{2h}}{(2n_h\widehat{\Pi}_h(1 - \widehat{\Pi}_h)}, \tag{3.45}$$

where n_{1h} is the number of subjects in study h who received success ratings from both raters, n_{2h} is the number who received one success and one failure rating, n_{3h} is the number who received failure ratings from both raters, and $n_h = n_{1h} + n_{2h} + n_{3h}$. An overall measure of agreement among the studies is estimated by computing a weighted average of the individual $\widehat{k}_h$, yielding

$$\widehat{k} = \frac{\sum_{h=1}^{N} n_h\widehat{\Pi}_h(1 - \widehat{\Pi}_h)\widehat{k}_h}{\sum_{h=1}^{N} n_h\widehat{\Pi}_h(1 - \widehat{\Pi}_h)}, \tag{3.46}$$

To test $H_o : k_1 = k_2 = ...k_N$, Donner et al. proposed a goodness of fit test using the statistic

$$\chi_G^2 = \sum_{h=1}^{N}\sum_{l=1}^{3} \frac{\{n_{lh} - n_h\widehat{\pi}_{lh}(\widehat{k}_h)}{n_h\widehat{\pi}_{lh}(\widehat{k}_h)}, \tag{3.47}$$

where $\widehat{\pi}_{lh}(\widehat{k}_h)$ is obtained by replacing Π_h by $\widehat{\Pi}_h$ and k_h by $\widehat{k}$ in $\pi_{lh}(k_h)$; $l = 1, 2, 3$; $h = 1, 2, ..., N$. χ^2_G follows an approximate chi-square distribution with $N - 1$ degrees of freedom, under the null hypothesis. (see Donner and Klair, 1996 for detail).

Donner et al. (1996) also proposed another method of testing $H_o : k_1 = k_2 = ...k_N$ using a large sample variance approach. The estimated large sample variance of $\widehat{k_h}$ (Bloch and Kraemer 1989), Fleiss and Davies 1982) is given by

$$\widehat{Var}(\widehat{k_h}) = \frac{1 - \widehat{k_h}}{n_h}\{(1 - \widehat{k_h})(1 - 2\widehat{k_h}) + \frac{\widehat{k_h}(2 - \widehat{k_h})}{2\widehat{\Pi}_h(1 - \widehat{\Pi}_h)}\}. \tag{3.48}$$

Let

$$\widehat{W}_h = \frac{1}{\widehat{Var}(\widehat{k_h})}$$

and

$$\widetilde{k} = \sum_{h=1}^{N}(\widehat{W}_h \widehat{k}_h) / \sum_{h=1}^{N}(\widehat{W}_h),$$

an approximate test of H_o is obtained by referring

$$\chi^2_v = \sum_{h=1}^{N} \widehat{W}_h(\widehat{k}_h - \widetilde{k})^2$$

to the chi-square distribution with $N - 1$ degrees of freedom. The statistic χ^2_v is undefined if $\widehat{k}_h = 1$ for any h. Unfortunately, this event can occur with fairly high frequency in samples of small to moderate size. In contrast the goodness of fit statistic, χ^2_G, can be calculated except in the extreme boundary case of $\widehat{k}_h = 1$ for all $h = 1, 2, ..., N$, when a formal test of significance has no practical value. Neither test statistic can be computed when $\widehat{\Pi}_h = 0$ or 1 for any h, since then $\widehat{k}_h$ is undefined. Based on Monte Carlo study, the authors found that the two statistic have similar properties for large samples ($n_h > 100$ for all h). But for small sample sizes, clearly the goodness of fit statistic χ^2_G is preferable.

3.1.4 τ *statistic*

Jolayemi (1986, 1990) proposed a statistic for agreement measure, that uses the chi-square distribution. The statistic was initiated from the background of the R^2, the coefficient of determination, which is an index for the explained variability of a regression model, which was then extended to the square contingency table. The author proposed a theorem which was also proved that "Consider a I×I (square) contingency table obtained by classifying the same N subjects into one of possible I outcomes by two raters. Then the Pearson Chi-square (X^2) statistic for independence is at most $(I - 1)N$. That is $0 \leq X^2 \leq (I - 1)N$." see Jolayemi (1990) for the proof of this theorem. He then proposed a statistic for the measure of agreement, denoted by

$$\tau = \sqrt{\lambda}, \quad -1 < \tau < 1, \tag{3.49}$$

where λ, which is an R^2-type statistic (Jolayemi, 1986) is defined as

$$\lambda = \frac{X^2}{max(X^2)} \tag{3.50}$$

and max(X^2) has been proved to be $(I-1)N$, see Jolayemi, (1990); Adejumo et al. (2001). Thus,

$$\lambda = \frac{X^2}{(I-1)N} \tag{3.51}$$

The advantage this statistic is having over Kappa is that by the nature of ($\lambda = \tau^2$) one may make inference on τ also through λ which estimates the explained variability exhibited by the configuration of the table as done in regression analysis. The author also proposed some arbitrary division on the range of $\mid \tau \mid$ with the respective strength of agreement, as Landis and Koch (1977a) has also proposed for Cohen kappa statistic in Table 3.3, as in Table 3.6 below: And

Table 3.6: Range of $\mid \tau \mid$ statistic with the respective strength of agreement.

$\mid \tau \mid$ *statistic*	*Strength of agreement*
0.00-0.20	poor
0.21-0.40	slight
0.41-0.60	moderate
0.61-0.80	substantial
0.81-1.00	almost perfect.

when $\tau < 0$ the agreement is negative.

3.1.5 *Tetrachoric correlation coefficient*

As stated by Banerjee et al. (1999), there are situations where two raters use different thresholds due to differences in their visual perception or decision attitude. By "threshold" we mean the value along the underlying continuum above which raters regard abnormality as present. Furthermore, with such data, the probability of misclassifying a case across the threshold is clearly dependent on the true value of the underlying continuous variable; the more extreme the value (the further away from a specified threshold), the smaller the probability of misclassification. Since this is so for all the raters, their misclassification probabilities cannot be independent. Therefore, kappa-type measures (weighted and unweighted kappas, intraclass kappa) are inappropriate in such situations.

In a situation where the diagnosis is regarded as the dichotomization of an underlying continuous variable that is unidimensional with a standard normal distribution, the Tetrachoric Correlation Coefficient (TCC) is an obvious choice for estimating interrater agreement. TCC estimates specifically, the correlation between the actual latent (un-observable) variables characterizing the raters' probability of abnormal diagnosis, and is based on assuming bivariate normality of the raters' latent variables. Therefore, not only does the context under which TCC is appropriate differ from that for kappa-type measures, but quantitatively they estimate two different, albeit related entities (Kraemer 1997). Several twin studies have used the TCC as a statistical measure of concordance among monozygotic and dizygotic twins, with respect to certain dichotomized traits.

The TCC is obtained as the maximum likelihood estimate for the correlation coefficient in the bivariate normal distribution,when only information in the contingency table is available (Tallis 1962, Hamdan 1970).

The concordance correlation coefficients (CCC) was also proposed by Lin (1989, 1992) for measuring agreement when the variable of interest is continuous and this agreement index was defined in the context of comparing two fixed observers. Lin (2000), King and Chinchilli (2001), Barnhart and Williamson (2001) and Barnhart et al. (2002) later proposed another modified index of Lin (1989,1992) that can take care of the multiple fixed observers or raters when the rating scale is continuous. Also see Chinchinlli et al. (1996) for intraclass correlation coefficients for interobserver reliability measure.

3.1.6 *Weighted least squares (WLS) method for correlated kappa*

Barnhart and Williamson (2002) proposed an approach of testing the equality of two different kappa statistics using weighted least squares (WLS) by Koch et al. (1977) in order to determine the correlation between the two kappa statistics for valid inference. Assuming there are four categorical readings Y_{11}, Y_{12}, Y_{21} and Y_{22} assessed on the same sets of N subjects. The first two readings (Y_{11} and Y_{12}) are obtained under one condition or method and the last two readings (Y_{21} and Y_{22}) are also from the other condition or method. There are two different readings obtained from two different raters (to assess interrater agreement) or replicated readings by one rater (to assess intrarater agreement). Barnhart and Williamson were able to compare these two agreement values to determine whether or not the reproducibility between the two readings differs from method to method as well as observing the correlation of the two agreement values.

Barnhart and Williamson (2002) were interested in testing the hypothesis of equality of the two kappa statistics $\widehat{k}_1$ (from method 1) and $\widehat{k}_2$ (from method 2) obtained from the two bivariate marginal tables, that is, contingency table $Y_{11} \times Y_{12}$ and table $Y_{21} \times Y_{22}$ with cell counts (collapsed cell probabilities) y_{ij++} (π_{ij++}) and y_{++kl} (π_{++kl}) respectively. Each of these kappa statistics are obtained using the Appendix 1 of Koch et al. (1977), which presented k as an explicit function of Π called the response function, in the the following form

$$k = F(\Pi) \equiv \exp(A_4)\log(A_3)\exp(A_2)\log(A_1)A_0\Pi, \qquad (3.52)$$

where $\Pi = (\Pi_{1111}, \Pi_{1112}, \ldots, \Pi_{111j}, \Pi_{2111}, \ldots, \Pi_{jj12}, \ldots, \Pi_{jjjj})$ denote the $J^4 \times 1$ vector of the cell probabilities for $Y_{11} \times Y_{12} \times Y_{21} \times Y_{22}$ contingency table; and A_0, A_1, A_2, A_3 and A_4 are matrices defined in Appendix and the exponentiation and logarithm are taken with respect to everything on the right hand side of (3.52). The weighted least squares estimator for k is

$$\widehat{k} = F(\Pi) \equiv \exp(A_4)\log(A_3)\exp(A_2)\log(A_1)A_0P, \tag{3.53}$$

where P is the vector of the cell proportions of the $J \times J \times J \times J$ table, which estimate Π. Therefore the

$$\widehat{Cov}(\widehat{k}) = \left(\frac{\partial F}{\partial P}\right) V \left(\frac{\partial F}{\partial P}\right)' \tag{3.54}$$

where $V = (diag(P) - PP')/N$ is the estimated covariance matrix for P and

$$\frac{\partial F}{\partial P} = diag(B_4)A_4diag(B_3)^{-1}A_3diag(B_2)A_2diag(B_1)^{-1}A_1A_0. \tag{3.55}$$

where $B_1 = A_1A_0P$, $B_2 = \exp(A_2)\log(B_1)$, $B_3 = A_3B_2$ and $B_4 = \exp(A_4)\log(B_3)$.

Using (3.53) and (3.54), construct a Wald test for the hypothesis ($H_o : k_1 = k_2$) by using $Z - score$

$$Z = \frac{\widehat{k}_1 - \widehat{k}_2}{\sqrt{(var(\widehat{k}_1) + var(\widehat{k}_2) - 2Cov(\widehat{k}_1, \widehat{k}_2))}} \tag{3.56}$$

However, to compute (3.53) and (3.54) they used the matrices A_0, A_1, A_2, A_3 and A_4 in the Appendix to obtain various kappa indices. See Appendix A.6 for different matrices expressions for Cohen's kappa, Weighted kappa and Intraclass kappa as expressed by Barnhart and Williamson (2002).

3.2 *Modelling in Agreement measure*

Due to the wide disagreement about the usefulness of kappa statistic to assess rater agreement and rather than using a single number to summarize agreement, some authors have proposed modelling of the structure of agreement using log-linear and latent class models.

3.2.1 Loglinear models

Tanner and Young (1985a) proposed a modelling structure of agreement for nominal scales, by considering loglinear models to express agreement in terms of components, such as chance agreement and beyond chance agreement. Using the loglinear model approach one can display patterns of agreement among several observers, or compare patterns of agreement when subjects are stratified by values of a covariate. Assuming there are n subjects who are related by the same k raters ($k \geq 2$) into I nominal categories, they express chance agreement, or statistical independence of the ratings, using the following loglinear model representation:

$$log(m_{ij...l}) = \mu + \lambda_i^{R_1} + \lambda_j^{R_2} + ... + \lambda_l^{R_k}, \qquad i,j,...l = 1,2,...I \tag{3.57}$$

where $m_{ij...l}$ is the expected cell count in the $ij..l$th cell of the joint k-dimensional cross-classification of the ratings, μ is the overall effect, $\lambda_i^{R_k}$ is the effect due to categorization by the kth rater in the cth category ($k = 1,...,K; c = 1,...,I$), and $\sum_{i=1}^{I}\lambda_i^{R_1} = ... = \sum_{l=1}^{I}\lambda_l^{R_k} = 0$. A useful generalization of the independence model incorporates agreement beyond chance in the following fashion:

$$log(m_{ij...l}) = \mu + \lambda_i^{R_1} + \lambda_j^{R_2} + ... + \lambda_l^{R_k} + \delta_{ij...l}, \qquad i,j,...l = 1,2,...I \tag{3.58}$$

The additional term $\delta_{ij...l}$ represents agreement beyond chance for the $ij...l$th cell. To test a given hypothesis concerning the agreement structure, the parameters corresponding to the agreement component $\delta_{ij...l}$ are assigned to specific cells or groups of cells in the contingency table. The term $\delta_{ij...l}$ can be defined according to what type of agreement pattern is being investigated. For example, to investigate homogeneous agreement among $K = 2$ raters, one would define δ_{ij} to be equal to

$$\delta_{ij} = \begin{cases} \delta & \text{If} \quad i = j, \\ 0 & \text{if} \quad i \neq j, \end{cases} \tag{3.59}$$

On the other hand to investigate a possibly nonhomogeneous pattern of agreement, that is differential agreement by response category, one would consider $\delta_{ij} = \delta_i I(i = j),\ i,j = 1,2,...,I$, where the indicator $I(i = j)$ is defined as

$$I(i = j) = \begin{cases} 1 & \text{If} \quad i = j, \\ 0 & \text{if} \quad i \neq j, \end{cases} \tag{3.60}$$

This approach addresses the higher-order agreement ($k \geq 2$) as well as pairwise agreement (Tanner and Young 1985a). The parameters then describe conditional agreement. For instance, agreement between two raters for fixed ratings by the other raters. The major advantage of this method is that it allows one to model the structure of agreement rather than simply describing it with a single summary measure. Graham (1995) extended Tanner and Young's approach to accommodate one or more categorical covariates in assessing agreement pattern between two raters. The baseline for studying covariate effects is taken as the conditional independence model, thus allowing covariate effects on agreement to be studied independently of each other and of covariate effects on the marginal observer distributions. For example, the baseline model for two raters and a categorical covariate X is given by

$$\begin{aligned} log(m_{ij...l}) &= \mu + \lambda_i^{R_1} + \lambda_j^{R_2} + \lambda_x^X + \lambda_{ix}^{R_1 X} + \lambda_{jx}^{R_2 X}, \\ & \quad i, j, ...l = 1, 2, ...I \end{aligned} \tag{3.61}$$

where $\lambda_i^{R_1}$,$\lambda_j^{R_2}$ are as defined in equation (3.57), λ_x^X is the effect of the xth level of the covariate X, and $\lambda_{ix}^{R_k X}$ $(k = 1, 2)$ is the effect of the partial association between the kth rater and the covariate X. Given the level of the covariate X, the above model assumes independence between the two raters' reports. Graham uses this conditional independence model as the baseline from which to gauge the strength of agreement. The beyond-chance agreement is modelled as follows:

$$\begin{aligned} log(m_{ijx}) &= \mu + \lambda_i^{R_1} + \lambda_j^{R_2} + \lambda_x^X + \lambda_{ix}^{R_1 X} + \lambda_{jx}^{R_2 X} \\ & \quad + \delta^{R_1 R_2} I(i = j) + \delta_x^{R_1 R_2 X} I(i = j), \\ & \quad i, j, ...l = 1, 2, ...I \end{aligned} \tag{3.62}$$

where $\delta^{R_1 R_2} I(i = j)$ represents overall beyond-chance agreement, and $\delta_x^{R_1 R_2 X} I(i = j)$ represents additional chance-corrected agreement associated with the xth level of the covariate X. Agresti (1988) and Tanner and Young (1985b) proposed methods of modelling loglinear model for agreement and disagreement pattern in an ordinal scale respectively. Magnitude as well as the direction of disagreement in ordinal scale ratings is very important. The major advantage of loglinear model framework over kappa liked statistics is that it provides natural way of modelling "how" the chance-corrected frequencies differ across the off-diagonal bands of the cross classification table. Agresti (1988) proposed a model of agreement plus linear-by-linear association, which is the combination of the model of Tanner and Young (1985a) and the uniform association model of Goodman (1979) for bivariate cross-classifications of ordinal variables. The model is

$$log(m_{ij}) = \mu + \lambda_i^{R_1} + \lambda_j^{R_2} + \beta u_i v_j + \delta_{ij}, \quad i, j = 1, 2, ...I \tag{3.63}$$

where

$$\delta_{ij} = \begin{cases} \delta & \text{If} \quad i = j, \\ 0 & \text{if} \quad i \neq j, \end{cases} \tag{3.64}$$

$u_1 < u_2 < ... < u_I$ or $v_1 < v_2 < ... < v_I$ are fixed scores assigned to the response categories, and μ, $\lambda_i^{R_1}$, $\lambda_j^{R_2}$ and $m_i j$ are as defined in equation (3.57).

3.2.2 Latent-class models

Latent-class models were also proposed by several authors to investigate inter-rater agreement (Aickin 1990, Uebersax and Grove 1990, Agresti 1992, Agresti and lang (1993),Williamson and Manatunga (1997), Banerjee et al. (1999)).

These models express the joint distribution of ratings as a mixture of distributions for classes of an unobserved (latent) variable. Each distribution in the mixture applies to a cluster of subjects representing a separate class of a categorical latent variable, those subjects being homogeneous in some sense. Agresti (1992) described a basic latent-class model for interrater agreement data by treating both the observed scale and the latent variable as discrete. Latent class models focus less on agreement between the raters than on the agreement of each rater with the "true" rating.

For instance, suppose there are three different raters, namely R_1, R_2, R_3, who rate each of n subjects into I categories. The latent-class model assumes that there is an unobserved categorical scale X, with L categories, such that subjects in each category of X are homogeneous. Given the level of X and base on this homogeneous, the joint ratings of R_1, R_2 and R_3 are assumed to be statistically independent. This is referred to as local independence. For a randomly selected subject, let π_{ijlk} denote the probability of ratings (i, j, l) by raters (R_1, R_2, R_3) and categorization in class k of X. Also let m_{ijlk} be the expected frequencies for the R_1-R_2-R_3-X cross-classification. The observed data then constitute a three-way marginal table of an unobserved four-way table. The latent-class model corresponding to loglinear model (R_1X, R_2X, R_3X) is the nonlinear model, which fit can be used to estimate conditional probabilities of obtaining various ratings by the raters, given the latent class, is of the form

$$\begin{aligned} log(m_{ijl+}) &= \mu + \lambda_i^{R_1} + \lambda_j^{R_2} + \lambda_l^{R_3} + log\{\sum_{k=1}^{L} \exp(\lambda_k^X \\ &\quad + \lambda_{ik}^{R_1X} + \lambda_{jk}^{R_2X} + \lambda_{lk}^{R_3X})\}, \\ &\quad i, j, ...l = 1, 2, ...I \end{aligned} \quad (3.65)$$

In addition, estimates of the probabilities of membership in various latent classes, conditional on a particular pattern of observed ratings, and use these to make predictions about the latent class to which a particular subject belongs. It seems, therefore the combination of loglinear and latent-class modelling should be a useful strategy for studying agreement.

To fit latent-class models, one can use data augmentation techniques, such as the EM algorithms. The E (expectation) step of the algorithm approximates counts in the complete R_1-R_2-R_3-X table using the observed R_1-R_2-R_3 counts and the working conditional distribution of X, given the observed ratings. The M (maximization) step treats those approximate counts as data in the standard iterative reweighted least-squares algorithm for fitting loglinear models. Alternatively, following Haberman (1988), one could adopt for the entire analysis a scoring algorithm for fitting nonlinear models or a similar method for fitting loglinear models with missing data.

In the case of ordinal scale, latent-class models that utilize the ordinality of ordered categories (Bartholomew 1983) have also been applied to studies of rater agreement. Agresti and Lang (1993) also proposed a model that treats the unobserved variable X as ordinal, and assumes a linear-by-linear association between each classification and X, using scores for the observed scale as well as for the latent classes. The model is of the form

$$\begin{aligned} log(m_{ijlk}) &= \mu + \lambda_i^{R_1} + \lambda_j^{R_2} + \lambda_l^{R_3} + \lambda_k^X + \beta^{R_1X}\lambda_i X_k \\ &\quad + \beta^{R_2X}\lambda_j X_k + \beta^{R_3X}\lambda_l X_k, \\ &\quad i,j,l = 1,2,...I \end{aligned} \tag{3.66}$$

where $k = 1, 2, ..., L$, the categories for X as defined before.

Qu et al. (1992,1995) proposed another approach that posit an underlying continuous variable. so that instead of assuming a fixed set of classes for which local independence applies, one could assume local independence at each level of a continuous latent variable. Williamson and Manatunga (1997) extended Qu et al. (1995) latent-variable models to analyze ordinal-scale ratings, with I categories, arising from n subjects who are being assessed by K raters using D different rating methods. Williamson and Manatunga (1997) obtained overall agreement and subject-level agreement between the raters based on the marginal and association parameter, using the generalized estimating equations approach which allows for subject- and/ or rater-specific covariates to be included in the model (Liag and Zeger 1986). This proposed approach can also be utilized for obtaining estimates of intrarater correlations if the raters assesses the same sample on two or more occasions, assuming enough time has passed between the ratings to insure that the rater does not remember his or her previous ratings, Banerjee et al. (1999).

More on agreement modelling shall be presented in the future work by considering some selected models that were originally designed for square tables which can be used for modelling the ratings of a given number of subjects by two or more raters. Also, we shall try to consider negative binomial as a substitute to Poisson model when the resulted cross-classified table of ratings is sparse.

4 Empirical examples

In this section we present some working examples on measurement of both association and agreement with some of the statistics reviewed in this paper that could handled I×I contingency tables when $I > 2$. We selected the data in such a way that both association and agreement are measured under nominal and ordinal categorical scales. For nominal scales data, we used Goodman and Kruskal's τ and U coefficient for association while for ordinal scales data, we used γ coefficient, Somers' d coefficient and Kendal tau-b coefficient. However, in the case of agreement irrespective of the categorical scale we used Cohen kappa statistic k_c and Intraclass kappa statistic k_{In}.

4.1 Example 1.

Consider the data on journal citation among four statistical theory and methods journals during 1987-1989 (Stigler, 1994; Agresti, 1996). The more often that articles in a particular journal are cited, the more prestige that journal accrues. For citations involving a pair of journals X and Y, view it as a "victory" for X if it is cited by Y and a "defeat" for X if it is cites Y. The categories used are BIOM=*Biometrika* , COMM=*Communications in Statistics*, JASA=*Journal of the American Statistical Association*, JRSSB=*Journal of the Royal Statistical Society Series B*.

Table 4.1: Cross-classification table of cited journal and citing journal of four statistical theory and methods journals.

Category	Cited journal				
Citing journal	BIOM	COMM	JASA	JRSSB	Total
BIOM	714	33	320	284	1351
COMM	730	425	513	276	1944
JASA	498	68	1072	325	1963
JRSSB	221	17	142	188	568
Total	2163	543	2047	1073	5826

$$\begin{aligned} Goodman\ and\ Kruskal's\ \tau &= 0.07514195 \\ U &= 0.1878702 \\ Cohen\ kappa &= 0.2119863 \\ Intraclass\ kappa &= 0.1889034. \end{aligned}$$

4.2 Example 2.

Consider the data obtained by two pathologists that assessed 27 patients twice for the presence (Y) or absence (N) of dysplasia as presented by Baker et al. (1991) and Barnhart et. al (2002). The categories used are $1 = NN$ (dysplasia absence on both times), $2 = NY$ (dysplasia absence in the first time but presence in the second time), $3 = YN$ (dysplasia presence in the first time but absence in the second time), $4 = YY$ (dysplasia presence on both times).

Table 4.2: Cross-classification table of dysplasia assessment for 27 patients.

Category	Pathologist 2				
Pathologist 1	1	2	3	4	Total
1	9	4	1	6	20
2	0	1	0	0	1
3	0	0	0	0	0
4	1	1	0	4	6
Total	10	6	1	10	27

$$\begin{aligned}\gamma &= 0.5000\\ Somers'\ d &= 0.7409\\ Kendal\ tau-b &= 0.9617\\ Cohen\ kappa &= 0.2419006\\ Intraclass\ kappa &= 0.1789474.\end{aligned}$$

4.3 Example 3.

Consider the following data taken from Agresti (1990). Letters A, B, C, D, and E are the categorical scales used for the classification of the subjects by the two raters.

Table 4.3: Cross-classification of 100 items by two raters

Category	Rater 2					
Rater 1	A	B	C	D	E	Total
A	4	16	0	0	0	20
B	0	4	16	0	0	20
C	0	0	4	16	0	20
D	0	0	0	4	16	20
E	16	0	0	0	4	20
Total	20	20	20	20	20	100

$$\begin{aligned}Goodman\ and\ Kruskal's\ \tau &= 0.6\\ U &= 0.689082\\ Cohen\ kappa &= -3.46944e-017\\ Intraclass\ kappa &= -3.469447e-017.\end{aligned}$$

5 Summary results and Conclusion

5.1 Summary results

We present the summary of the examples results in the previous section for association and agreement measure.

Table 5.1: Summary table of results from the five examples on association and agreement.

	Association					Agreement by Kappa	
Example	τ	U	γ	$Somers'd$	$K.tau-b$	$Cohen$	$Intraclass$
1	0.0751	0.1879				0.2120	0.1889
2			0.5000	0.7409	0.9617	0.2419	0.1789
3	0.6000	0.6891				$-3.469E-017$	$-3.469E-017$

From these results in Table 5.1, we observed that both association and agreement may have very low values as we have in example 1. Association may be very high while agreement will be small as we have in example 2. Also, as in example 3 there may be very strong association without any strong agreement because the agreement value is less than zero. We used the a pair (path B and path E) of the data given by Holmquist et al. (1967) that investigated the variability in the classification of carcinoma in situ of the uterine cervix on 118 slides by seven pathologists and we had very high values for the two measures. However, a similar result with example 2 was recorded when we also used another data on multiple sclerosis assessment as presented by (Basu et al. 1999). More analyzes can be done for agreement under modelling as we have earlier mentioned in § 3.2 but we shall reserve these for the future work on modelling some special models for agreement and some others.

5.2 Conclusion

We have already showed that measures of association and agreement statistics are different from one another based on all the literatures presented in this paper. We presented up till date measures under each of the measurements. And we observed from the results of the working examples that agreement is a subset of association, that is, agreement can be regarded as a special case of association. When there is a strong or low agreement between two raters or observers, strong or low association will also exits between them. However, there may be strong association without any strong or low agreement, this can occur if one rater consistently rates subjects one or more levels higher than the other rater, then there will be a strong association between them, but the strength of agreement will be very weak. Once there is an agreement between two raters irrespective of the strength, or level, association will definitely exists also, but strong association can exists with no strong agreement. Hence agreement can be regarded as a subset of association.

6 References

[**1**] Adejumo, A. O., Sanni, O. O. M. and Jolayemi, E. T. (2001): Imputation Method for Two Categorical Variables. *Journal of the Nigerian Statisticians*, 4(1), 39-45.

[**2**] Agresti, A. (1988). A model for agreement between ratings on an ordinal scales, *Biometrics*, 44, 539- 548.

[**3**] Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statist. Methods Med. Res.*, 1, 201-218.

[**4**] Agresti, A. (1996). *An introduction to Categorical Data Analysis*, Wiley, New York.

[**5**] Agresti, A. and Lang, J. B. (1993). Quasi-symmetry latent class models, with application to rater agreement. *Biometrics*,49, 131-139.

[**6**] Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive model, and it relation to Cohen's kappa. *Biometrics*, 46, 293-302.

[**7**] Baker, S. G., Freedman, L. S., and Parmar, M. K. B. (1991). Using replicate observations in observer agrrement studies with binary assessments. *Biometrics* 47, 1327-1338.

[**8**] Banerjee, M., Capozzoli, M., Mcsweeney, L., and Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measure. *The Cana. J. of. Statist.*, 27(1), 03-23.

[**9**] Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics*, 52, 695- 702.

[**10**] Barlow, W., Lai, M. Y., and Azen, S. P (1991). A comparison of methods for calculating a stratified kappa. *Statist. Med.*, 10, 1465-1472.

[**11**] Barnhart, H. X, and Williamson, J. M. (2002). Weighted least-squares approach for comparing correlated kappa. *Biometrics*, 58,1012-1019.

[**12**] Barnhart, H. X, Michael, H., and Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, 58,1020-1027.

[**13**] Bartholomew, D. J. (1983). Latent variable models for ordered categorical data. *J. Econometrics*, 22, 229-243.

[**14**] Basu, S., Basu, A., and Raychaudhuri, A. (1999). Measuring agreement between two raters for ordinal response: a model based approach. *The Statistician*, 48(3), 339-348.

[**15**] Bergsma, W. P. (1997). *Marginal Models for Categorical data*. University Press, Tilburg, The Netherland.

[**16**] Block, D. A., and Kraemar, H. C. (1989). 2 x 2 kappa coefficients: Measures of agreement or association. *Biometrics*, 45, 269-287.

[**17**] Chinchilli, V. M., Martel, J. K., Kumanyika, S., and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, 52, 341-353.

[**18**] Cicchetti, D. V. (1972). A new measure of agreement between rank ordered variables. *Proceedings, 80th Annual convention, Americ. Psych. Associ.*, 17-18.

[**19**] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Edu. and Psych. Meas.*, 20, 37-46.

[**20**] Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psych. Bull.*, 70, 213-220.

[**21**] Darroch, J. N., and McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian J. Statist.*, 28, 371-388.

[**22**] Davies, M. and Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, 38, 1047-1051.

[**23**] Donner, A., and Eliasziw, M. (1992). A goodness of fit approach to inference procedures for the kappa statistic: confidence interval construction, significance testing and sample size estimation. *Statist. Med.*, 11, 1511-1519.

[**24**] Donner, A., and Eliasziw, M. (1997). A hierarchical approach to inference concerning interobserver agreement for multinomial data. *Statist. Med.*, 16, 1097-1106.

[**25**] Donner, A. and Klair, N. (1996). The statistical analysis of kappa statistics in multiple samples. *J. Clin. Epidemiol.*, 49, 1053-1058.

[**26**] Donner, A., Eliasziw, M. and Klar, N. (1996). Testing homogeneity of kappa statistics. *Biometrics*, 52, 176-183.

[**27**] Everitt, B.S. (1968). Moments of the statistics kappa and weighted kappa. *British J. Math. Statist. Psych.*, 21, 97-103.

[**28**] Forthofer, R. N., and Koch, G. G. (1973). An analysis for compounded functions of categorical data. *Biometrics*, 29, 143-157.

[**29**] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psych. Bull.*, 76, 378-382.

[**30**] Fleiss, J. L. (1973). *Statistical methods for rates and proportions.* Wiley, New York, 144-147.

[**31**] Fleiss, J. L. and Cicchetti, D. V. (1978). Inference about weighted kappa in the no-null case. *Appl. Psych. Meas.*, 2, 113-117.

[**32**] Fleiss, J. L. and Cohen J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures reliability. *Educ. and Psych. Meas.*, 33, 613-619.

[**33**] Fleiss, J. L. and Davies, M. (1982). Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *Amer. J. Epidemiol.*, 115, 841-845.

[**34**] Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa *Psych. Bull.*, 72, 323-327.

[**35**] Gonin, R., Lipsitz, S. R., Fitzmaurice, G. M., and Molenberghs, G. (2000). Regression modelling of weighted k by using generalized estimating equations. *Appl. Statist.*, 49, 1-18.

[**36**] Goodman L. A. (1979). Simple models for the analysis of association in cross classifications having ordered categories. *J. Amer. Statist. Assoc.*, 74, 537-552.

[**37**] Goodman L. A. and Kruskal, W. H. (1954). Measuring of association for cross classifications. *J. Amer. Statist. Assoc.*, 49, 732-768.

[**38**] Graham, P. (1995). Modeling covariate effects in observer agreement studies: The case of nominal scale agreement. *Statist. Med.*, 14, 299-310.

[**39**] Grover, R. and Srinnivasan, V. (1987). General surveys on beverages, *J. Marketing Research*, 24: 139-153.

[**40**] Guggenmoos-Holzmann, I. (1993). How reliable are chance-corrected measures of agreement? *Statist. Med.*, 12(23): 2191-2205.

[**41**] Haberman, S. J. (1988). A stabilized Newton-raphson algorithm for loglinear models for frequency tables derived by indirect observation *Sociol. Methodol.*, 18, 193-211.

[**42**] Hamdan, M. A. (1970). The equivalence of tetrachoric and maximum likelihood estimates of ρ in $2{\times}2$ tables. *Biometrika*, 57, 212-215.

[**43**] Holmquist, N. S., McMahon, C. A. and Williams, O. D. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology*, 84, 334-345.

[**44**] Jolayemi , E. T. (1986). Adjust R^2 method as applied to loglinear models. *J. Nig. Statist. Assoc.*, 3, 1-7.

[**45**] Jolayemi , E. T. (1990). On the measure of agreement between two raters. *Biometrika*, 32(1), 87-93.

[**46**] Kendall, M. G. (1945). The treatment of ties in rank problems. *Biometrika*, 33, 239-251.

[**47**] Kendall, M. G. and Stuart, A. (1961). *The advance theory of statistics.* Vol. 2. Hafner Publication Company, New York.

[**48**] Kendall, M. G. and Stuart, A. (1979). *The advance theory of statistics.* Vol. 2.Inference and Relationship 4th edition, Macmillian, New York.

[**49**] Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. Jr., and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, 33, 133-158.

[**50**] Kozlowski, S.W.J.,and Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *J. Applied Psycho.*, 77(2), 161-167.

[**51**] Kraemer, H. C. (1997). What is the "right" statistical measure of twin concordance for diagnostic reliability and validity? *Arch. Gen. Psychiatry*, 54, 1121-1124.

[**52**] Kraemer, H. C., Bloch, D. A. (1988). Kappa coefficients in epidemiology: an appraisal of a reappraisal. *J. Clin. Epidem.*, 41, 959-968.

[**53**] Landis, J. R, and Koch, G. G. (1975a). A review of statistical methods in the analysis of data arising from observer reliability studies (Part I) . *Statistica Neerlandica*, 29, 101-123.

[**54**] Landis, J. R, and Koch, G. G. (1975b). A review of statistical methods in the analysis of data arising from observer reliability studies (Part II) . *Statistica Neerlandica*, 29, 151-161.

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

[**56**] Landis, J. R., and Koch, G. G. (1977b). A one-way components of variance model for categorical data. *Biometrics*, 33, 159-174.

[**57**] Liag, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

[**58**] Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268.

[**59**] Maclure, M., and Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *Amer. J. Epidemiol.*, 126(2), 161- 169.

[**60**] Maclure, M., and Willett, W. C. (1988). Misinterpretation and misuse of the kappa statistic. (Dissenting letter and reply) *Amer. J. Epidemiol.*, 128(5), 1179-1181.

[**61**] Oden, N. L. (1991). Estimating kappa from binocular data. *Statist. Med.*, 10,1303-1311.

[**62**] O'Connell, D. L., and Dobson, A. J. (1984). General obsrver-agreement measures on individual subjects and groups of subjects. *Biometrics*, 40, 973-983.

[**63**] Posner, K. L., Sampson, P. D., Caplan, R. a., Ward, R. J., and Cheney, F. w. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statist. Med.*, 9, 1103-1115.

[**64**] Qu, Y., Williams, G.W., Beck, G. J., and Medendorp, S. V. (1992). Latent variable models for clustered dichotomous data with multiple subclusters. *Biometrics*, 48, 1095-1102.

[**65**] Qu, Y., Piedmonte, M. R., and Medendorp, S. V. (1995). Latent variable models for clustered ordinal data. *Biometrics*, 51, 268-275.

[**66**] Schouten, H. J. A. (1993). Estimating kappa from binocular data and comparing marginal probabilities. *Statist. Med.*, 12, 2207-2217.

[**67**] Shoukri, M. M. (2004). *Measures of interobserver agreement*. Chapman and Hall.

[68] Shoukri, M. M., Martin, S. W., and Mian, I. U. H. (1995). Maximum likelihood estimation of the kappa coefficient from models of matched binary responses. *Statist. Med.* 14, 83-99.

[69] Somer, R. H. (1962). A new asymmetric measure of association ordinal variables. *Ameri. Sociol. Review*, 27, 799-811.

[70] Stigler, S. M. (1994). Citation patterns in the journal of statistics and probability. *Statist. Sci.*, 9, 94-108.

[71] Tallis, G. M. (1962). The maximum likelihood estimation of correlations from contingency tables. *Biometrics*, 18, 342-353.

[73] Tanner, M. A., and Young, M. A. (1985a). Modeling agreement among raters. *J. Amer. Statist. Assoc.*, 80, 175-180.

[74] Tanner, M. A., and Young, M. A. (1985b). Modeling ordinal scale agreement . *Psych. Bull.*, 98, 408-415.

[72] Theil, H. (1970). On the estimation of relationships involving qualitative variables. *Ameri. J. Sociol.*, 76, 103-154.

[75] Thompson, W. D. and Walter, S. D. (1988a). A reappraisal of the kappa coefficient. *J. Clini. Epidemiol.*, 41(10), 949-958.

[76] Thompson, W. D. and Walter, S. D. (1988b). Kappa and the concept of independent errors. *J. Clini. Epidemiol.*, 41(10), 969-970.

[77] Uebersax, J. S., and Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statist. Med.*, 9, 559- 572.

[78] Williamson, J. M., and Manatunga, A. K. (1997). Assessing interrater agreement from dependent data. *Biometric*, 53, 707-714

[79] Yule, G. U. (1900). On the association of attributes in statistics. *Phil. Trans.*, Ser. A 194, 257-319.

[80] Yule, G. U. (1912). On the methods of measuring association between two attributes. *J. Roy. Statist. Soc.*, 75, 579-642.

A Appendix

Most of the statistics reviewed under measures of association as well as agreement can be expressed in $exp-log$ notation (Forthofer and Koch, 1973; Bergsma, 1997 and Barnhart and Williamson (2002). Consider the fraction $(\pi_1+\pi_2)/(\pi_3+$

π_4). In matrix notation, this expression is

$$\begin{aligned}
\frac{\pi_1+\pi_2}{\pi_3+\pi_4} &= \exp\left[\log(\pi_1+\pi_2)-\log(\pi_3+\pi_4)\right] \\
&= \exp\left[\begin{pmatrix} 1 & -1 \end{pmatrix} \log\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix}\right].
\end{aligned}$$

In general, any product of strictly positive terms involves exponentiating the sum of the logarithms of the terms.

A.1 Gamma statistic

$$\gamma = \frac{\pi_c - \pi_d}{\pi_c + \pi_d} \tag{A.1}$$

So Gamma in $\exp - \log$ format is

$$\begin{aligned}
\gamma &= \frac{\pi_c - \pi_d}{\pi_c + \pi_d} \\
&= \frac{\pi_c}{\pi_c + \pi_d} - \frac{\pi_d}{\pi_c + \pi_d} \\
&= \exp\{\pi_c - \log(\pi_c + \pi_d)\} - \exp\{\pi_d - \log(\pi_c + \pi_d)\} \\
&= \begin{pmatrix} 1 & -1 \end{pmatrix} \exp\left[\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \log\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \pi_c \\ \pi_d \end{pmatrix}\right].
\end{aligned}$$

A.2 Somer's-d statistic

$$\Delta_{BA} = \frac{\pi_c - \pi_d}{1 - \pi_{t,A}} \tag{A.2}$$

So Δ_{BA} in the " $\exp - \log$ " notation is as follows:

$$\Delta_{BA} = \begin{pmatrix} 1 & -1 \end{pmatrix} \exp\left[\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \log\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}\begin{pmatrix} \pi_c \\ \pi_d \\ 1'\pi \\ \pi_{t,A} \end{pmatrix}\right]$$

where $1'\pi = \sum_i \pi_i = 1$, (this is done so that a function of π is obtained; $''1''$ is not a function of π.

A.3 Kendall's tau-b

$$\tau_b = \frac{\pi_c - \pi_d}{\sqrt{(1-\pi_{t,A})(1-\pi_{t,B})}} \tag{A.3}$$

This statistic can also be written in $\exp - \log$ way as

$$\exp\left[\begin{pmatrix} 1 & 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \log \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \pi_c \\ \pi_d \\ 1'\pi \\ \pi_{t,A} \\ \pi_{t,B} \end{pmatrix}\right]$$

A.4 Pearson's correlation coefficient

$$\rho = \frac{cov(A,B)}{\sigma_A \sigma_B} = \frac{E(AB) - E(A)E(B)}{\sigma_A \sigma_B} \tag{A.4}$$

This statistic in the $\exp - \log$ notation, ρ is written as

$$\rho_{A,B} = \begin{pmatrix} 1 & -1 \end{pmatrix} \exp\left[\begin{pmatrix} 0 & 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 1 & 1 & 0 & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \log \begin{pmatrix} E(A) \\ E(B) \\ E(AB) \\ \sigma_A^2 \\ \sigma_B^2 \end{pmatrix}\right]$$

The variances of A and B can be written as

$$\begin{aligned} \begin{pmatrix} \sigma_A^2 \\ \sigma_B^2 \end{pmatrix} &= \begin{pmatrix} E(A^2) - (E(A))^2 \\ E(B^2) - (E(B))^2 \end{pmatrix} \\ &= \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \times \\ & \qquad \exp\left[\begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \log \begin{pmatrix} E(A) \\ E(A^2) \\ E(B) \\ E(B^2) \end{pmatrix}\right]. \end{aligned}$$

Let π_{ij} be the cell probability for cell (i,j). The $E(A) = \sum_i a_i \pi_{i+}$ and $E(B) = \sum_j b_j \pi_{+j}$, where a_i and b_j are scores of categories I of A and J of B respectively. Let M_r and M_c be such that $M_r'\pi$ and $M_r'\pi$ produce the row and column totals respectively. Let a and a^2 be the vectors with elements a_i and a_i^2 respectively. Also, let D_{ab} be the diagonal matrix with element $a_i b_j$ on the main diagonal. Then the expected values that are used are

$$\begin{pmatrix} E(A) \\ E(A^2) \\ E(B) \\ E(B^2) \\ E(AB) \end{pmatrix} = \begin{pmatrix} \sum_i a_i \pi_{i+} \\ \sum_i a_i^2 \pi_{i+} \\ \sum_j b_j \pi_{+j} \\ \sum_j b_j^2 \pi_{+j} \\ \sum_{ij} a_i b_j \pi_{ij} \end{pmatrix} = \begin{pmatrix} a' M_r' \\ a^{2\prime} M_r' \\ b' M_c' \\ b^{2\prime} M_c' \\ 1' D_{ab}' \end{pmatrix} \pi.$$

Therefore, ρ is a sum of products of sums of products of sums of probabilities.

A.5 Cohen's kappa

For 2×2 contingency table, Cohen's kappa in § 3.1.1 can be expressed in $\exp - \log$ notation and also to illustrate the matrix notation of matrices A_0, A_1, A_2, A_3 and A_4 mentioned in § 3.1.6 above Barnhart and Williamson (2002). Cohen' kappa was given as

$$\widehat{k} = \frac{\sum_{i=1}^{I} \pi_{ii} - \sum_{i=1}^{I} \pi_{i.}\pi_{.i}}{1 - \sum_{i=1}^{I} \pi_{i.}\pi_{.i}} \tag{A.5}$$

For 2×2 contingency table, this can be written as

$$
\begin{aligned}
k &= \frac{\sum_{i=1}^{2} \pi_{ii} - \sum_{i=1}^{2} \pi_{i.}\pi_{.i}}{1 - \sum_{i=1}^{2} \pi_{i.}\pi_{.i}} \\
&= \frac{(\pi_{11} + \pi_{22}) - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})}{1 - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})} \\
&= \exp\left[\log\left\{(\pi_{11} + \pi_{22}) - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})\right\} - \log\left\{1 - (\pi_{1+}\pi_{+1} + \pi_{2+}\pi_{+2})\right\}\right] \\
&= \exp\begin{pmatrix} 1 & -1 \end{pmatrix} \log\begin{pmatrix} -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & 1 \end{pmatrix}\begin{pmatrix} \pi_{1+}\pi_{+1} \\ \pi_{2+}\pi_{+2} \\ \pi_{11} + \pi_{22} \\ 1 \end{pmatrix} \\
&= \exp\begin{pmatrix} 1 & -1 \end{pmatrix} \log\begin{pmatrix} -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & 1 \end{pmatrix} \exp\begin{pmatrix} \log(\pi_{1+}\pi_{+1}) \\ \log(\pi_{2+}\pi_{+2}) \\ \log(\pi_{11} + \pi_{22}) \\ \log(1) \end{pmatrix} \\
&= \exp\begin{pmatrix} 1 & -1 \end{pmatrix} \log\begin{pmatrix} -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & 1 \end{pmatrix} \exp\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\
&\quad \times \log\begin{pmatrix} \pi_{1+} \\ \pi_{2+} \\ \pi_{+1} \\ \pi_{+2} \\ \pi_{11} + \pi_{22} \\ 1 \end{pmatrix} \\
&= \exp\begin{pmatrix} 1 & -1 \end{pmatrix} \log\begin{pmatrix} -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & 1 \end{pmatrix} \exp\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\
&\quad \times \log\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{pmatrix} \\
&= \exp(A_4) log(A_3) \exp(A_2) \log(A_1)\Pi. \qquad \text{(A.6)}
\end{aligned}
$$

Matrix A_1 produces a vector with the row marginal, column marginal, diagonal sum, and the total sum of the cell probabilities.
Matrix A_2 produces a vector with four main quantities in the log scale of k.
Matrix A_3 produces a vector of the numerator and denominator of k; and
Matrix A_4 divides the numerator by the denominator to produce k.
This is just for a single kappa statistic using Cohen, this can also be done for other kappa indices (Landis and Koch, 1977a).

A.6 Response function F for various kappa indices

We present response function F in equation (3.52) for various kappa indices when we need to estimate two different kappa statistics for Π based on the two methods or conditions under consideration (Barnhart and Williamson (2002)). Firstly, we present the general formulae for the matrices in equation (3.52).

$$\begin{aligned} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} &= F(\Pi) = \exp(A_4) log(A_3) \exp(A_2) \log(A_1) A_0 \Pi \\ &= \exp \begin{pmatrix} A_{44} & 0 \\ 0 & A_{44} \end{pmatrix} \log \begin{pmatrix} A_{33} & 0 \\ 0 & A_{33} \end{pmatrix} \\ &\quad \times \exp \begin{pmatrix} A_{22} & 0 \\ 0 & A_{22} \end{pmatrix} \log \begin{pmatrix} A_{11} & 0 \\ 0 & A_{11} \end{pmatrix} A_0 \Pi, \end{aligned}$$

where A_0 is a $2J^2 \times J^4$ matrix of the form

$$A_0 = \begin{pmatrix} e'_{J^2} & 0 & \dots & 0 \\ 0 & e'_{J^2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & e'_{J^2} \\ I_{J^2} & I_{J^2} & \dots & I_{J^2} \end{pmatrix}$$

and e_J is a $J \times 1$ vector of all ones, I_J is the $J \times J$ identity matrix with dimension J, 0 is a matrix of all zeros with dimensions conforming to the other part of the block matrices.

For each of the kappa indices, we have the following:

1. *Cohen's kappa coefficient*:

$$A_{44} = \begin{pmatrix} 1 & -1 \end{pmatrix},$$

$$A_{33} = \begin{pmatrix} -e'_J & 1 & 0 \\ -e'_J & 0 & 1 \end{pmatrix},$$

$$A_{22} = \begin{pmatrix} I_J & I_J & 0 \\ 0 & 0 & I_2 \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} e'_J & 0 & \dots & 0 \\ \vdots & & \vdots & \\ 0 & 0 & \dots & e'_J \\ I_J & I_J & \dots & I_J \\ e'_J & e'_J & \dots & e'_J \end{pmatrix},$$

where A_{44} is 1×2 matrix, A_{33} is $2 \times (J+2)$, A_{22} is $(J+2) \times (2J+2)$, A_{11} is $(2J+2) \times J^2$ and $I_J(j)$ is the jth row of the identity matrix I_J.

2. *Weighted kappa coefficient*:

$$A_{44} = \begin{pmatrix} 1 & -1 \end{pmatrix},$$

$$A_{33} = \begin{pmatrix} -w' & 1 & 0 \\ -w' & 0 & 1 \end{pmatrix},$$

$$A_{22} = \begin{pmatrix} e_J & 0 & \dots & 0 & I_J & 0 \\ 0 & e_J & \dots & 0 & I_J & 0 \\ \vdots & \vdots & \vdots & \vdots & & \\ 0 & 0 & \dots & e_J & I_J & 0 \\ 0 & 0 & \dots & 0 & 0 & I_2 \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} e'_J & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & e'_J \\ I_J & I_J & \dots & I_J \\ & & W' & \\ e'_J & e'_J & \dots & e'_J \end{pmatrix},$$

where $W = (w_{11}, w_{12}, \dots, w_{JJ})'$ is $J^2 \times 1$ vector of weights. A_{33} is $2 \times (J^2 + 2)$ matrix, A_{22} is $(J^2 + 2 \times (2J + 2)$, A_{11} is $(2J + 2) \times J^2$ and A_{44} is as defined above.

3. *Intraclass kappa coefficient*: Using equation (3.35) we have,

$$A_{44} = \begin{pmatrix} 1 & -1 \end{pmatrix},$$

$$A_{33} = \begin{pmatrix} -e'_J & 1 & 0 \\ -e'_J & 0 & 1 \end{pmatrix},$$

$$A_{22} = \begin{pmatrix} 2I_J & 0 \\ 0 & I_2 \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} \frac{e'_J + I_J(1)}{2} & \frac{I_J(1)}{2} & \dots & \frac{I_J(1)}{2} \\ \frac{I_J(2)}{2} & \frac{e'_J + I_J(2)}{2} & \dots & \frac{I_J(2)}{2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{I_J(J)}{2} & \frac{I_J(J)}{2} & \dots & \frac{e'_J + I_J(J)}{2} \\ I_J(1) & I_J(2) & \dots & I_J(J) \\ e'_J & e'_J & \dots & e'_J \end{pmatrix},$$

where A_{22} is $(J + 2) \times (J + 2)$ matrix, A_{11} is $(J + 2) \times J^2$, A_{33} and A_{44} are as defined above.