STATISTICA, anno LXXVII, n. 4, 2017

MODELING ASSOCIATION PLUS AGREEMENT AMONG MULTI-RATERS FOR ORDERED CATEGORIES

Ayfer Ezgi Yilmaz [1]
Department of Statistics, Faculty of Science, Hacettepe University, Beytepe-Ankara, Turkey

1. INTRODUCTION

Square contingency tables are frequently used in many fields, such as medicine, sociology, and behavioral sciences. Several inter-rater reliability coefficients have been proposed in the literature. The Cohen (1960) kappa coefficient is the most used agreement index. Kappa coefficients provide useful information about the reliability of data. Numerous extensions and generalizations of kappa coefficient have been proposed in the literature. Depending on the scale type (nominal, ordinal, and interval) and the number raters, different coefficients should be used. For ordinal categories, weighted kappa coefficient was suggested for use (Cohen, 1968). As alternatives to weighted kappa coefficient, weighted version of Bangdiwala (1988) B and Gwet (2012) $AC2$ coefficients were also suggested. There has been also suggested measure of agreement $OR(Ag)$ coefficient based on the odds ratio (Attanasio *et al.*, 2010).

For the multi-rater studies, Light (1971) κ which is the generalized form of Cohen's κ, Fleiss (1971) κ, and Hubert (1977) κ can be used. von Eye and Mun (2005) adapted the raw agreement, kappa coefficient, Brennan and Prediger (1981) κ_η coefficient for multi-rater studies. Berry *et al.* (2007, 2008) suggested kappa coefficients for nominal and ordinal square tables with multi-raters. Hubert's κ coefficient was also reformulated for the ordinal tables (Warrens, 2005). Quatto (2004) proposed a procedure for testing chance agreement among multiple raters.

Kappa coefficients are always applicable, easy to calculate and interpret, available in general purpose statistical software packages, and they condense relevant information into one coefficient. Besides, it is possible to summarize the rater agreement with a single number. However, most authors have criticized it because of the limitations and insufficiencies, such as: loss of information, unless κ approaches 1, the measure does not allow one to describe the structure of the joint frequency distribution, specific hypotheses cannot be tested, and covariates cannot be taken into account (Kundel and Polansky,

[1] Corresponding Author. E-mail: ezgiyilmaz@hacettepe.edu.tr

2003; Tanner and Young, 1985). Kappa coefficients have been criticized because of their dependency on the rater prevalence (Kottner *et al.*, 2011). Feinstein and Cicchetti (1990) and Cicchetti and Feinstein (1990) made two well-known paradoxes with Cohen's κ: (1) a low kappa can occur at a high agreement, and (2) unbalanced marginal distributions produce higher values of kappa than more balanced marginal distributions.

Because of the insufficiency of kappa coefficients, most authors prefer to use log-linear models. Instead of summarizing agreement, log-linear models analyze the structure of the agreement in the data (Tanner and Young, 1985). The log-linear model studies give more detailed information about the data. There are specialized log-linear models for each different scale type (nominal, ordinal, and interval) and the type of model changes according to the number raters.

In recent studies, researchers pay more attention to the assessment of more than two rater's agreement instead of two. Although there is a huge literature on inter-rater reliability coefficients and the agreement models for two raters, the models are not sufficient for the multi-rater studies. The existing models in multi-rater case are sometimes insufficient to explain the structure of the data. In this article, we first present a review about the inter-rater reliability coefficients and log-linear association plus agreement models. Then, we propose different modifications to analyze the association plus partial or global agreement. These models allow the distinguishability between adjacent categories to vary according to their positions and allows to investigate the agreement as global or as partial. For more clarification, we use two medical data sets to illustrate the suggested modifications.

The inter-rater reliability coefficients are reviewed in Section 2. The log-linear models for two and multi rater studies are reviewed in Section 3. Section 4 presents the suggested log-linear models to analyze the association plus partial or global agreement. Section 5 presents the illustrative examples, followed by conclusion in Section 6.

2. INTER-RATER RELIABILITY COEFFICIENTS

Cohen's weighted kappa coefficient was suggested for use in two rater studies with ordered categories. For the multi-rater studies, Light (1971), Hubert (1977), and Berry *et al.* (2007) weighted kappa coefficients can be used.

Let n_{ij} denote the number of objects, π_{ij} denote the probability of cell (i, j), and n denote the total number of observations. Let $\pi_{i.}$ indicate the ith row total probability and $\pi_{.j}$ indicate the jth column total probability in a $R \times R$ contingency table. The weighted kappa coefficient κ_w is

$$\kappa_w = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}\pi_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}\pi_{i.}\pi_{.j}}{1 - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}\pi_{i.}\pi_{.j}}, \tag{1}$$

where w_{ij} are the weights with $0 \leq w_{ij} \leq 1$. The values of weights have been discussed in many studies. The most popular weights for weighted kappa are the linear and the

quadratic weights (Cicchetti and Allison, 1971; Fleiss and Cohen, 1973).

Light's kappa coefficient is the arithmetic mean of all possible pairs of the raters. Instead of using unweighted kappa coefficient, it is possible to calculate the weighted version of Light's κ coefficient using of weighted kappas. Let h be the number of raters and $\kappa_w(ij)$ be the weighted kappa coefficient among ith and jth raters. Lights's $L.\kappa_w$ coefficient is given in Equation (2).

$$L.\kappa_w = \frac{2}{h(h-1)} \sum_{i=1}^{h-1} \sum_{j=i+1}^{h} \kappa_w(ij). \tag{2}$$

Suppose there are three raters and R is the number of categories. Let p_i, q_j, and r_k be the marginal proportions and $A = \{a_{ij}\}$, $B = \{b_{ij}\}$, and $C = \{c_{ij}\}$ be the sub-tables. The calculation of sub-tables and the marginal probabilities are given in Equation (3) and (4), respectively.

$$a_{ij} = \sum_{k=1}^{R} \pi_{ijk} \quad b_{ij} = \sum_{k=1}^{R} \pi_{jik} \quad c_{ij} = \sum_{k=1}^{R} \pi_{jki}, \tag{3}$$

and

$$p_i = \sum_{j=1}^{R} \sum_{k=1}^{R} \pi_{ijk} \quad q_i = \sum_{j=1}^{R} \sum_{k=1}^{R} \pi_{jik} \quad r_i = \sum_{j=1}^{R} \sum_{k=1}^{R} \pi_{jki}. \tag{4}$$

The observed agreement P_0^H and the proportion agreement expected by chance P_e^H of Hubert's weighted kappa coefficient are defined as

$$P_0^H = \frac{1}{3} \sum_{i=1}^{R} \sum_{j=1}^{R} \left[1 - \frac{|i-j|}{R-1} \right] (a_{ij} + b_{ij} + c_{ij}), \tag{5}$$

and

$$P_e^H = \frac{1}{3} \sum_{i=1}^{R} \sum_{j=1}^{R} \left[1 - \frac{|i-j|}{R-1} \right] (p_i q_j + p_i r_j + q_i r_j). \tag{6}$$

Then, Hubert's weighted kappa coefficient is $H.\kappa_w = \frac{P_0^H - P_e^H}{1 - P_e^H}$.

Berry *et al.* (2007, 2008) suggested a weighted kappa coefficient for three rater studies with ordinal categories. P_0^M and P_e^M are

$$P_0^M = \sum_{i=1}^{R} \sum_{j=1}^{R} \sum_{k=1}^{R} w_{ijk} \pi_{ijk}, \tag{7}$$

and

$$P_e^M = \sum_{i=1}^{R}\sum_{j=1}^{R}\sum_{k=1}^{R} w_{ijk}\, p_i q_j r_k. \tag{8}$$

Then Mielke, Berry, and Johnston's weighted kappa coefficient is $M.\kappa_w = \frac{P_0^M - P_e^M}{1-P_e^M}$. Here the weights w_{ijk} are calculated from Equation (9).

$$w_{ijk} = 1 - \frac{|i-j| + |i-k| + |j-k|}{2(R-1)}. \tag{9}$$

3. LOG-LINEAR MODELS

Although it is possible to summarize the inter-rater reliability by the coefficients, most authors prefer to use log-linear agreement models. Agreement models are suggested to apply on the square contingency tables with nominal categories. The single way to apply agreement models to the ordered tables is to ignore the ordered structure of the variables. However, this will lead loss of information. In order to analyze the agreement of ordered square contingency tables, the association models with agreement parameter which analyze agreement and association together are suggested. In addition to analysis of agreement, odds ratios may be calculated under fitted.

3.1. Log-linear models for two raters

Consider an $R \times R$ contingency table where the first rater is represented by X and the second rater is represented by Y. In this two-way table, cross-classifies are multinominal sample of n subjects on two categorical responses. The independence, agreement, uniform association (UA), and uniform association plus agreement (UAA) models are summarized in Table 1.

TABLE 1
Log-linear models for $R \times R$ contingency tables.

Model	Variable*	Equation	Df^+	Ref.
Independence	N,O	$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$	$(R-1)^2$	Agresti (1984)
Agreement	N	$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \delta_{ij}$	$(R-1)^2 - 1$	Tanner and Young (1985)
UA	O	$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta \times u_i v_j$	$R^2 - 2R$	Goodman (1979)
UAA	O	$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta \times u_i v_j + \delta_{ij}$	$R^2 - 2R - 1$	Agresti (1988)

* N: Nominal; O: Ordinal

+ *df*: Degrees of freedom

In Table 1, λ is an overall effect parameter, λ_i^X is the effect of variable X at i and λ_j^Y is the effect of variable Y at j with the constraints $\sum_{i=1}^{R} \lambda_i^X = \sum_{j=1}^{R} \lambda_j^Y = 0$. δ_{ij} in the agreement model is the agreement parameter which is given in Equation (10).

$$\delta_{ij} = \begin{cases} \delta & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

In the association model, β is the association parameter, $u_i = i$ and $v_j = j$ are the estimated score values. The UAA model has been extensively used to describe both agreement and association among the raters.

To describe the variation of distinguishability between adjacent categories, non-uniform association model (NUA) was suggested by Valet *et al.* (2007). Differently from the uniform association model, this model includes $(R-1)$ association parameters $\beta_{k,k+1}$. $\beta_{k,k+1}$ is the association parameter among the categories k and $k+1$. NUA model allows for the calculation of different values of odds ratios on the main diagonal. The corresponding log-linear model is as given in Equation (11).

$$\log\, m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y - \frac{|i-j|}{2} \times \sum_{k=\min(i,j)}^{\max(i,j)-1} \beta_{k,k+1}. \tag{11}$$

The non-uniform association plus agreement model (NUAA) which is used to describe both agreement and non-uniform association among raters is also discussed (Valet *et al.*, 2007). NUAA model can be written as

$$\log\, m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y - \frac{|i-j|}{2} \times \sum_{k=\min(i,j)}^{\max(i,j)-1} \beta_{k,k+1} + \delta_{ij}, \tag{12}$$

where $i, j = 1, 2, ..., R$. The agreement parameter is as defined in Equation (10). NUA model has $df = R^2 - 3R + 2$ and NUAA model has one more parameter than NUA model.

Association between the variables of a square contingency table can be simply expressed using odds ratios. The odds ratio for adjacent categories is

$$\theta_{i,i+1} = \frac{m_{i,i} \times m_{i+1,i+1}}{m_{i+1,i} \times m_{i,i+1}}, \tag{13}$$

where $i = 1, 2, ..., (R-1)$.

3.2. Association plus agreement models for multi-raters

Let X, Y, and Z be the raters of an $R \times R \times R$ table which have ordered categories. The independence model (M0), agreement model (M1) (Tanner and Young, 1985), uniform association model (M2) (Agresti, 1984), and uniform association plus agreement models with different combinations of association and agreement parameters (M3-M7) have been discussed for multi-raters (Melia and Diener-West, 1994; Lawal, 2003). These models are summarized in Table 2 for three-rater studies.

TABLE 2
The independent, agreement, uniform association, and uniform association plus agreement models for three-rater tables.

Model	Equation
M0	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$
M1	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk}$
M2	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 \times u_i v_j + \beta_2 \times u_i w_k + \beta_3 \times v_j w_k + \beta_4 \times u_i v_j w_k$
M3	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 \times u_i v_j + \beta_2 \times u_i w_k + \beta_3 \times v_j w_k + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk}$
M4	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 \times u_i v_j + \beta_2 \times u_i w_k + \beta_3 \times v_j w_k + \delta_{ij} + \delta_{ik} + \delta_{jk}$
M5	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 \times u_i v_j + \beta_2 \times u_i w_k + \beta_3 \times v_j w_k + \delta_{ijk}$
M6	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 \times u_i v_j + \beta_2 \times u_i w_k + \beta_3 \times v_j w_k + \beta_4 \times u_i v_j w_k + \delta_{ij} + \delta_{ik} + \delta_{jk}$
M7	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 \times u_i v_j + \beta_2 \times u_i w_k + \beta_3 \times v_j w_k + \beta_4 \times u_i v_j w_k + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk}$

In Table 2, $u_i = i$, $v_j = j$, and $w_k = k$ are the score values for X, Y, and Z, respectively. β_1 is the association parameter among X–Y, β_2 is the association parameter among X–Z, and β_3 is the association parameter among Y–Z. Here δ_{ij}, δ_{ik}, and δ_{jk} are the partial agreement parameters that show the agreement among X–Y, X–Z, and Y–Z, respectively. δ_{ijk} is the global agreement parameter that shows the agreement among X, Y, and Z. The partial and global agreement parameters are given in Equations (14)–(17).

$$\delta_{ij} = \begin{cases} \delta & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \qquad (14)$$

$$\delta_{ik} = \begin{cases} \delta & \text{if } i = k, \\ 0 & \text{otherwise} \end{cases} \qquad (15)$$

$$\delta_{jk} = \begin{cases} \delta & \text{if } j = k, \\ 0 & \text{otherwise} \end{cases} \qquad (16)$$

$$\delta_{ijk} = \begin{cases} \delta & \text{if } i = j = k, \\ 0 & \text{otherwise} \end{cases} \qquad (17)$$

4. The suggested non-uniform association plus agreement models for multi-raters

Although there is a huge literature for two rater studies, there is not enough literature to explain the association and agreement among multi-raters. In this article, several modifications of non-uniform association plus agreement models are developed for multi-raters. Besides, instead of classical score values, a modification of score values are discussed.

4.1. Non-uniform association plus agreement models

Let X, Y, and Z be the raters with ordered categories ($R \geq 3$). δ_{ij}, δ_{ik}, and δ_{jk} are the agreement parameters defined as in Equations (14)–(17), respectively. Let $\beta_{l,l+1}$ be the

association parameter between the adjacent categories l and $(l+1)$ of X and Y, $\varphi_{l,l+1}$ be the association parameter between the adjacent categories l and $(l+1)$ of X and Z, $\omega_{l,l+1}$ be the association parameter between the adjacent categories l and $(l+1)$ of Y and Z where $l = 1,2,...(R-1)$. The association among all of the raters is called global association. ε is the global association parameter which shows the association among X, Y, and Z. Here the global association parameter is weighted by the weights as in Equation (9). Differently from the non-uniform association model for two raters, here the non-uniform association parameters are weighted by Cicchetti and Allison (1971) linear weights. The suggested non-uniform association and non-uniform association plus agreement models for three-raters are summarized in Table 3.

TABLE 3

The non-uniform association and non-uniform association plus agreement models for three-raters.

Model	Equation and *df*
M8	$\log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{R-1} \times \sum_{l=\min(i,j)}^{\max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{R-1} \times \sum_{l=\min(i,k)}^{\max(i,k)-1} \varphi_{l,l+1} - \frac{\lvert j-k\rvert}{R-1} \times \sum_{l=\min(j,k)}^{\max(j,k)-1} \omega_{l,l+1}$ $df = R^3 - 6R + 5$
M9	$\log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{R-1} \times \sum_{l=\min(i,j)}^{\max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{R-1} \times \sum_{l=\min(i,k)}^{\max(i,k)-1} \varphi_{l,l+1} - \frac{\lvert j-k\rvert}{R-1} \times \sum_{l=\min(j,k)}^{\max(j,k)-1} \omega_{l,l+1} + \delta_{ij} + \delta_{ik} + \delta_{jk}$ $df = R^3 - 6R + 2$
M10	$\log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{R-1} \times \sum_{l=\min(i,j)}^{\max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{R-1} \times \sum_{l=\min(i,k)}^{\max(i,k)-1} \varphi_{l,l+1} - \frac{\lvert j-k\rvert}{R-1} \times \sum_{l=\min(j,k)}^{\max(j,k)-1} \omega_{l,l+1} + \delta_{ijk}$ $df = R^3 - 6R + 4$
M11	$\log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{R-1} \times \sum_{l=\min(i,j)}^{\max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{R-1} \times \sum_{l=\min(i,k)}^{\max(i,k)-1} \varphi_{l,l+1} - \frac{\lvert j-k\rvert}{R-1} \times \sum_{l=\min(j,k)}^{\max(j,k)-1} \omega_{l,l+1} + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk}$ $df = R^3 - 6R + 1$
M12	$\log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{R-1} \times \sum_{l=\min(i,j)}^{\max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{R-1} \times \sum_{l=\min(i,k)}^{\max(i,k)-1} \varphi_{l,l+1} - \frac{\lvert j-k\rvert}{R-1} \times \sum_{l=\min(j,k)}^{\max(j,k)-1} \omega_{l,l+1} - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} \times \varepsilon$ $df = R^3 - 6R + 4$
M13	$\log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{R-1} \times \sum_{l=\min(i,j)}^{\max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{R-1} \times \sum_{l=\min(i,k)}^{\max(i,k)-1} \varphi_{l,l+1} - \frac{\lvert j-k\rvert}{R-1} \times \sum_{l=\min(j,k)}^{\max(j,k)-1} \omega_{l,l+1} - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} \times \varepsilon + \delta_{ijk}$ $df = R^3 - 6R + 3$

4.2. Global association plus agreement models

As ε is the global association parameter which shows the association among the three raters (X, Y, and Z), the global association plus partial or global agreement models with

linear weights are summarized in Table 4.

TABLE 4
The global association plus agreement models for three-raters.

Model	Equation	*Df*
M14	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} \times \varepsilon + \delta_{ijk}$	R^3-3R
M15	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} \times \varepsilon + \delta_{ij} + \delta_{ik} + \delta_{jk}$	R^3-3R-2
M16	$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)} \times \varepsilon + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk}$	R^3-3R-3

5. ILLUSTRATIVE EXAMPLES

5.1. The uterine cervix data set

The data in Table 5 is based on the data originally discussed by Holmquist *et al.* (1967). This data set has also been analyzed in the studies of Landis and Koch (1977a), Becker and Agresti (1992), and Saracbasi (2011). To investigate the variability in the classification of carcinoma in situ of the uterine cervix, three pathologists classified 118 slides into the 5 categories. Because the data contained sampling zero frequencies, the categories were reclassified to the following categories: (1) Negative, (2) Atypical Squamous Hyperplasia, (3) Carcinoma in Situ + Squamous Carcinoma with Early Stromal Invasion + Invasive Carcinoma (Becker and Agresti, 1992; Landis and Koch, 1977a).

TABLE 5
Independent classification by three pathologists of most involved histological lesion.

		C		
A	B	1	2	3
1	1	18	4	0
	2	1	1	0
	3	0	2	0
2	1	2	3	0
	2	3	4	0
	3	4	10	0
3	1	0	0	0
	2	0	2	1
	3	3	16	44

The weighted kappa coefficients among the pathologists are calculated and the results are summarized in Table 6. Landis and Koch (1977b) interpretation levels of kappa are added to Table 6. The results show that there is a "substantial" agreement among

the three pathologists. The agreement among B–C is less than the agreement among the other pairs.

TABLE 6
The results of weighted kappa coefficients.

Pathologists	Coefficient	Estimate	Level
A–B	κ_w	0.713	Substantial
A–C	κ_w	0.615	Substantial
B–C	κ_w	0.497	Moderate
A–B–C	$L.\kappa_w$	0.606	Substantial
A–B–C	$H.\kappa_w$	0.605	Substantial
A–B–C	$M.\kappa_w$	0.608	Substantial

Table 7 shows the goodness-of-fit test results and the parameter estimates of M0–M16 when $G^2 = \sum_{ij} n_{ij} \log(n_{ij}/m_{ij})$ is the likelihood ratio-statistic (Sokal and Rohlf, 1981). According to the presented results, all the models, except M0 and M1, fit the data sufficiently well. Akaike Information Criteria $[AIC = G^2 - 2\text{df}]$ and Bayesian Information Criteria $[BIC = G^2 - \ln(n)\text{df}]$ can be used to select the best fitting model.

TABLE 7
The results of goodness-of-fit test and the parameter estimates for the models.

Model	G^2	*Df*	P-value	AIC	BIC
M0	195.630	20	0.000	–	–
M1	45.697	16	0.000	–	–
M2	19.679	16	0.235	-12.321	-56.652
M3	14.830	13	0.318	-11.170	-47.189
M4	17.095	14	0.251	-10.905	-49.695
M5	15.936	16	0.457	**-16.064**	**-60.395**
M6	16.144	13	0.241	-9.856	-45.875
M7	13.877	12	0.309	-10.123	-43.371
M8	10.452	14	0.728	-17.548	-56.337
M9	5.693	11	0.893	-16.307	-46.785
M10	6.969	13	0.904	-19.031	-55.050
M11	5.267	10	0.873	-14.733	-42.440
M12	6.767	13	0.914	**-19.233**	**-55.252**
M13	6.734	12	0.875	-17.266	-50.514
M14	23.009	18	0.190	-12.991	**-62.863**
M15	19.238	16	0.256	-12.762	-57.093
M16	16.567	15	0.345	-13.433	-54.993

In order to discuss the structure of the models and to compare their results, the models are classified as uniform association models (M1-M7), non-uniform association models (M8-M13), and models with global association with linear weights (M14-M16). According to the represented results in Table 7, the best fitting model among the uniform association models is M5, among the non-uniform association models is M12, and

among the models with global association with linear weights is M14. The parameter estimates of each three models are summarized in Table 8.

M5 contains the partial uniform association and global agreement parameters. According to the parameter estimates of M5, the highest association is between pathologists A and B, and the lowest one is between B and C. There is an agreement ($\hat{\delta} > 0$) among the three pathologists. M12 contains the non-uniform association and global association parameters. According to the parameter estimates of M12, there is a high association between the three pathologists. There are positive associations ($\hat{\beta}_{12}, \hat{\beta}_{23} > 0$) between the adjacent categories of pathologists A and B. When there is a negative association ($\hat{\varphi}_{12} < 0$) between "negative" and "atypical squamous hyperplasia" of A and C, there is a positive and higher association ($\hat{\varphi}_{23} > 0$) between the "atypical squamous hyperplasia" and "carcinoma in situ + squamous carcinoma with early stromal invasion + invasive carcinoma".

TABLE 8
The parameter estimates of M5, M10, and M14 models.

Model	Parameter	Estimate	Std. Error	P-value
M5	β_1	1.390	0.391	0.000
	β_2	1.273	0.438	0.004
	β_3	0.331	0.339	0.330
	δ	0.885	0.417	0.034
M12	β_{12}	1.270	0.758	0.094
	β_{23}	0.329	0.897	0.714
	φ_{12}	-0.890	0.977	0.362
	φ_{23}	3.392	1.356	0.012
	ω_{12}	-0.020	0.770	0.980
	ω_{23}	0.277	1.110	0.803
	ε	2.808	1.496	0.061
M14	ε	4.313	0.885	0.000
	δ	-0.178	0.616	0.772

Within all the represented models, the best fitting one is M14 which contains global association and global agreement parameters. Based on the M14 results, there is a high association, but disagreement ($\hat{\delta} < 0$) among the three pathologists. The association plus agreement models are expressed in terms of the conditional odds ratios. For M5 model, the formulation of conditional log-odds ratios and calculated odds ratios are shown in the following matrices. In the following matrix, $l = 1$ for $\hat{\theta}_{ij(k)}$, $l = 2$ for $\hat{\theta}_{i(j)k}$, and $l = 3$ for $\hat{\theta}_{(i)jk}$.

$$\log\hat{\theta}_{M5}=\begin{bmatrix}\hat{\beta}_l+\hat{\delta} & \hat{\beta}_l\\ \hat{\beta}_l & \hat{\beta}_l\\ \hline \hat{\beta}_l+\hat{\delta} & \hat{\beta}_l-\hat{\delta}\\ \hat{\beta}_l-\delta & \hat{\beta}_l+\hat{\delta}\\ \hline \hat{\beta}_l & \hat{\beta}_l\\ \hat{\beta}_l & \hat{\beta}_l+\hat{\delta}\end{bmatrix}\quad \hat{\theta}_{ij(k)}=\begin{bmatrix}9.73 & 4.01\\ 4.01 & 4.01\\ \hline 9.73 & 1.66\\ 1.66 & 9.73\\ \hline 4.01 & 4.01\\ 4.01 & 9.73\end{bmatrix}\hat{\theta}_{i(j)k}=\begin{bmatrix}8.65 & 3.57\\ 3.57 & 3.57\\ \hline 8.65 & 1.47\\ 1.47 & 8.65\\ \hline 3.57 & 3.57\\ 3.57 & 8.65\end{bmatrix}\hat{\theta}_{(i)jk}=\begin{bmatrix}3.37 & 1.39\\ 1.39 & 1.39\\ \hline 3.37 & 0.57\\ 0.57 & 3.37\\ \hline 1.39 & 1.39\\ 1.39 & 3.37\end{bmatrix}$$

For M12 model, the formulation of conditional log-odds ratios and calculated odds ratios are shown in the following matrices. In the following matrix, $\hat{\alpha}=\hat{\beta}$ for $\hat{\theta}_{ij(k)}$, $\hat{\alpha}=\hat{\varphi}$ for $\hat{\theta}_{i(j)k}$ and $\hat{\alpha}=\hat{\omega}$ for $\hat{\theta}_{(i)jk}$. The odds ratios of the each category of a layer are equal, such as: $[\hat{\theta}_{ij(1)}=\hat{\theta}_{ij(2)}=\hat{\theta}_{ij(3)}]$.

$$\log\hat{\theta}_{M12}=\begin{bmatrix}\hat{\alpha}_{1,2}+\hat{\varepsilon}/2 & (\hat{\alpha}_{1,2}+\hat{\alpha}_{2,3})/2\\ (\hat{\alpha}_{1,2}+\hat{\alpha}_{2,3})/2 & \hat{\alpha}_{2,3}+\hat{\varepsilon}/2\end{bmatrix}$$

$$\hat{\theta}_{ij(k)}=\begin{bmatrix}14.49 & 2.22\\ 2.22 & 5.66\end{bmatrix}\hat{\theta}_{i(j)k}=\begin{bmatrix}1.67 & 3.49\\ 3.49 & 120.98\end{bmatrix}\hat{\theta}_{(i)jk}=\begin{bmatrix}3.99 & 1.14\\ 1.14 & 5.37\end{bmatrix}$$

For M14 model, the conditional odds ratios are equal. By the M14 model, the formulation of conditional log-odds ratios and calculated odds ratios are shown in the following matrices.

$$\hat{\theta}_{ijk}=\hat{\theta}_{ij(k)}=\hat{\theta}_{i(j)k}=\hat{\theta}_{(i)jk}.$$

$$\log\hat{\theta}_{M14}=\begin{bmatrix}\hat{\beta}/2+\hat{\delta} & 0\\ 0 & \hat{\beta}/2\\ \hline \hat{\beta}/2+\hat{\delta} & -\hat{\delta}\\ -\hat{\delta} & \hat{\beta}/2+\hat{\delta}\\ \hline \hat{\beta}/2 & 0\\ 0 & \hat{\beta}/2+\hat{\delta}\end{bmatrix}\quad \hat{\theta}_{ijk}=\begin{bmatrix}7.23 & 1.00\\ 1.00 & 8.64\\ \hline 7.23 & 1.19\\ 1.19 & 7.23\\ \hline 8.64 & 1.00\\ 1.00 & 7.23\end{bmatrix}$$

According to the odds ratios from the matrix above, when the pathologist A's decision is "negative" or "atypical squamous hyperplasia"; the odds of giving "atypical squamous hyperplasia" decision rather than "negative" decision of pathologist B is 7.23 times higher than giving "atypical squamous hyperplasia" decision rather than "negative" decision of pathologist C. When the pathologist A' decision is "carcinoma in situ + squamous carcinoma with early stromal invasion + invasive carcinoma", the odds of giving "atypical squamous hyperplasia" decision rather than "negative" decision of pathologist B is 8.64 times higher than giving "atypical squamous hyperplasia" decision rather than "negative" decision of pathologist C. Because odds ratios on main diagnosis diverge from 1, decisions of pathologist are more similar than one level up category of carcinoma in situ of uterine cervix. Thus, there is an agreement between their decisions.

5.2. *The liver metastases data set*

The data in Table 9 is taken from Uebersax (1992). In order to investigate the liver metastases, three tests were performed. The categories of liver metastases are: (1) Definitely negative results, (2) Marginal results, (3) Definitely positive results.

TABLE 9
Independent classification by three tests of liver metastases.

Tests				Tests				Tests			
T1	T2	T3	Obs.	T1	T2	T3	Obs.	T1	T2	T3	Obs.
1	1	1	36	2	2	1	12	3	1	3	1
1	1	2	22	2	2	2	25	3	2	1	1
1	2	1	26	2	2	3	5	3	2	2	7
1	2	2	22	2	3	1	1	3	2	3	10
1	3	1	3	2	3	2	5	3	3	1	3
1	3	3	1	2	3	3	10	3	3	2	13
2	1	1	13	3	1	1	1	3	3	3	66
2	1	2	14	3	1	2	1				

TABLE 10
The results of goodness-of-fit test and the parameter estimates for the models.

Model	G^2	*Df*	P-value	AIC	BIC
M0	400.050	20	0.000	–	–
M1	134.956	16	0.000	–	–
M2	32.732	16	0.008	–	–
M3	40.972	13	0.000	–	–
M4	51.171	14	0.000	–	–
M5	51.639	16	0.000	–	–
M6	32.215	13	0.002	–	–
M7	24.474	12	0.018	–	–
M8	19.758	14	0.138	-8.242	**-60.001**
M9	18.491	11	0.071	-3.509	-44.177
M10	19.758	13	0.101	-6.242	-54.304
M11	9.291	10	0.505	-10.709	-47.680
M12	19.188	13	0.117	-6.812	-54.874
M13	9.771	12	**0.636**	**-14.229**	-58.595
M14	43.558	18	0.001	–	–
M15	52.412	16	0.000	–	–
M16	41.621	15	0.000	–	–

Table 10 shows the goodness-of-fit test results and the parameter estimates of M0–M16. According to the presented results, only seven of seventeen models fit the data. The models with uniform association parameters do not fit the data. In that case, non-uniform association plus agreement models are good alternatives to uniform ones. Ac-

cording to BIC results in Table 10, the best fitting model is non-uniform association model (M8). Yet non-uniform association plus global agreement model (M13) is the best fitting model according to AIC and goodness-of-fit test results.

6. Conclusions

In recent studies, interrater agreement analysis has grown extensively. There are different ideas among the researchers when the subject is reliability or agreement. In practice, because they summarize the rater agreement with a single number, some researchers prefer using the kappa-like statistics. Some researchers criticize the kappa coefficients and assert to use log-linear models instead of them. The main argument of the researchers who prefer to use agreement models reveals pure agreement.

There is a huge literature for two-rater agrement studies. There are numerous measures of reliability and log-linear agreement models for each table structure and different number of raters. To get more reliable results, researchers would like to take the advises of more than two experts. When Landis and Koch (1977a) analyzed the uterine cervix data set, they used pairwise agreement statistics between the raters. Becker and Agresti (1992) applied log-linear models to the data among pairs of raters. Although one option is to investigate agreement as pairs, it is not possible to analyze the overall agreement or association.

Instead of investigating the agreement as pairs, Melia and Diener-West (1994) suggested the seven models to analyze the agreement and association together. As we illustrate in the liver metastases data set, these models are not always fit the data. In this article, we proposed some modifications of log-linear models for the study of interrater agreement in the case that has multi-raters with ordered categories. Because the existing models are insufficient and do not explain data well, we focus on the non-uniform and global associations, and also the partial and global agreement from the point of log-linear models.

Each of the seventeen models contain different parameters, such as: global association, partial uniform association, partial non-uniform association, partial agreement, and global agreement. The best fitting model changes depend on the data set. As a result of the analysis, all the models or more than one model can be found as statistically significant. In that case, information criteria can be helpful to select the best fitting model. Then, all interpretations can be made from the best fitting model. The conditional odds ratios vary between these models.

Different models can also be chosen depend on the aim of the study. If the aim is to interpret the associations as pairwise, M2-M7 can be useful. The proposed non-uniform models allows the distinguishability between the categories to vary. It is possible to interpret the association between the categories for each rater pair. For a more detailed association contraction between the categories of the variable, M8-M13 can be preferable to M2-M7. For a more detailed contraction of agreement or disagreement, the models with partial agreement parameters can be used. M7 (within the models of uniform asso-

ciation) and M11 (within the models of non-uniform association) give the most detailed information about the positive or negative association and agreement or disagreement of the raters.

In this article, we described the application of log-linear models for modeling agreement that overcome some of the limitations of the kappa-like statistics. Rather than summarizing agreement by a single number, log-linear models can be used the structure of the agreement in the data. Although the coefficients have been proposed to use of two or three raters, the log-linear models can be easily generalized for more than three raters. The models can be extended for the tables which subjects are stratified by a covariate.

Log-linear models help us to interpret the square tables with odds ratios which are calculated from expected values of best fitting model. To draw more reliable inferences, κ coefficient which is calculated from the expected values of best fitting model can be helpful to summarize the table with only one value.

References

A. Agresti (1984). *Analysis of Ordinal Categorical Data*. John Wiley & Sons, New York.

A. Agresti (1988). *A model for agreement between ratings on an ordinal scale*. Biometrics, 44, no. 2, pp. 539–548.

M. Attanasio, M. Enea, L. Rizzo (2010). *Some issues concerning the statistical evaluation of a screening test: The arfi ultrasound case*. Statistica, 70, no. 3, pp. 311–322.

S. I. Bangdiwala (1988). *The agreement chart*. In *Technical report*, University of North Carolina at Chapel Hill, Department of Biostatistics, Institute of Statistics Mimeo.

M. P. Becker, A. Agresti (1992). *Log-linear modelling of pairwise interobserver agreement on a categorical scale*. Statistics in Medicine, 11, no. 1, pp. 101–114.

P. W. Berry, K. J. Berry, J. E. Johnson (2007). *The exact variance of weighted kappa with multiple raters*. Psychological Reports, 101, pp. 655–660.

P. W. Berry, K. J. Berry, J. E. Johnson (2008). *Resampling probability values for weighted kappa with multiple raters*. Psychological Reports, 102, pp. 606–613.

R. L. Brennan, D. J. Prediger (1981). *Coefficient kappa: Some uses, misuses, and alternatives*. Educational and Psychological Measurement, 41, pp. 687–699.

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

D. V. Cicchetti, A. R. Feinstein (1990). *High agreement but low kappa: II. resolving the paradoxes*. Journal of Clinical Epidemiology, 43, no. 6, pp. 551–558.

J. COHEN (1960). *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, 20, no. 1, pp. 37–46.

J. COHEN (1968). *Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit*. Psychological Bulletin, 70, no. 4, pp. 213–220.

A. R. FEINSTEIN, D. V. CICCHETTI (1990). *High agreement but low kappa: I. the problems of the two paradoxes*. Journal of Clinical Epidemiology, 43, no. 6, pp. 543–549.

J. L. FLEISS (1971). *Measuring nominal scale agreement among many raters*. Psychological Bulletin, 76, no. 5, pp. 378–382.

J. L. FLEISS, J. COHEN (1973). *The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability*. Educational and Psychological Measurement, 33, pp. 613–619.

L. A. GOODMAN (1979). *Simple models for the analysis of association in cross-classifications having ordered categories*. Journal of the American Statistical Association, 74, no. 367, pp. 537–552.

K. L. GWET (2012). *Handbook of Inter-Rater Reliability*. Advanced Analytics, LLC, Maryland.

N. S. HOLMQUIST, C. A. MCMAHON, O. D. WILLIAMS (1967). *Variability in classification of carcinoma in situ of the uterine cervix*. Archives of Pathology, 84, pp. 334–345.

L. HUBERT (1977). *Kappa revisited*. Psychological Bulletin, 84, no. 2, pp. 289–297.

J. KOTTNER, L. AUDIGE, S. BRORSON, A. DONNER, B. J. GAJEWSKI, A. HROBJARTSSON, C. ROBERTS, M. SHOUKRI, D. L. STREINER (2011). *Guidelines for reporting reliability and agreement studies (GRRAS) were proposed*. Journal of Clinical Epidemiology, 64, pp. 96–106.

H. L. KUNDEL, M. POLANSKY (2003). *Measurement of observer agreement*. Radiology, 228, no. 2, pp. 303–308.

J. R. LANDIS, G. G. KOCH (1977a). *An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers*. Biometrics, 33, no. 2, pp. 363–374.

J. R. LANDIS, G. G. KOCH (1977b). *The measurement of observed agreement for categorical data*. Biometrics, 33, no. 1, pp. 159–174.

B. LAWAL (2003). *Categorical Data Analysis with SAS and SPSS Applications*. Lawrence Erlbaum Associates Inc, New Jersey.

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

B. M. MELIA, M. DIENER-WEST (1994). *Modeling Inter Rater Agreement for Pathologic Features of Choroidal Melanoma*. John Wiley & Sons, New York.

P. QUATTO (2004). *Testing agreement among multiple raters*. Statistica, 64, no. 1, pp. 145–151.

T. SARACBASI (2011). *Agreement models for multiraters*. Turkish Journal of Medical Science, 41, no. 5, pp. 939–944.

R. R. SOKAL, F. J. ROHLF (1981). *Biometry*. Freeman, New York.

M. A. TANNER, M. A. YOUNG (1985). *Modeling agreement among raters*. Journal of the American Statistical Association, 80, no. 389, pp. 175–180.

J. S. UEBERSAX (1992). *Modeling approaches for the analysis of observer agreement*. Investigative Radiology, 27, no. 9, pp. 738–743.

F. VALET, C. GUINOT, J. Y. MARY (2007). *Log-linear non-uniform association models for agreement between two ratings on an ordinal scale*. Statistics in Medicine, 26, pp. 647–662.

A. VON EYE, E. Y. MUN (2005). *Analyzing Rater Agreement: Manifest Variable Methods*. Lawrence Erlbaum Associates Inc, New Jersey.

M. J. WARRENS (2005). *Inequalities between multi-rater kappas*. Advanced in Data Analysis and Classification, 4, pp. 271–286.

SUMMARY

In square contingency tables, analysis of agreement between the row and column classifications is of interest. In such tables, the kappa-like statistics are used as a measure of reliability. In addition to the kappa coefficients, several authors discussed agreement in terms of log-linear models. Log-linear agreement models are suggested for use to summarize the degree of agreement between nominal variables. To analyze the agreement between ordinal categories, the association models with agreement parameter can be used. In the recent studies, researchers pay more attention to the assessment of agreement among more than two raters' decisions, especially in areas of medical and behavioral sciences. This article focuses on the approaches to study of uniform and non-uniform association with inter-rater agreement for multi-raters with ordered categories. In this article, we proposed different modifications of association plus agreement models and illustrate use of the approaches over two numerical examples.

Keywords: Global agreement; Partial agreement; Uniform association; Non-uniform association; Log-linear model; Ordinal scales.