Journal of Data Science 15(2017), 1-24

Assessing agreement between raters from the point of coefficients and log-linear models

Ayfer Ezgi Yilmaz[1*] , Tulay Saracbasi[1]

[1] *Department of Statistics, Hacettepe University*

Abstract: In square contingency tables, analysis of agreement between row and column classifications is of interest. For nominal categories, kappa co- efficient is used to summarize the degree of agreement between two raters. Numerous extensions and generalizations of kappa statistics have been pro- posed in the literature. In addition to the kappa coefficient, several authors use agreement in terms of log-linear models. This paper focuses on the approaches to study of inter-rater agreement for contingency tables with nominal or ordinal categories for multi-raters. In this article, we present a detailed overview of agreement studies and illustrate use of the approaches in the evaluation agreement over three numerical examples.

Key words: Agreement, kappa, log-linear models, multi-raters, nominal, or- dinal, weights.

1. Introduction

Square contingency tables are frequently used in many fields, such as medicine, sociology, and behavioral sciences. The R× R tables, in which classified variables are intimately related, are called square contingency tables. Square contingency tables may arise in different ways (Lawal, 2003):

- When a sample of individuals or subjects is cross-classified according to two essentially similar categorical variables.
- When samples of pairs of matched individuals or subjects such as husbands and wives, fathers and sons, or twin brothers are classified according to some categorical variable of interest.
- In panel studies where each individual or subject in a sample is classified according to the same criterion at two different points in time.
- In rating experiments in which a sample of N individuals or subjects is rated independently by the same two raters into one of R nominal or ordinal categories.

When working on these kinds of tables, firstly the agreement between row and column variables is investigated. Interrater agreement represents the extent to which different judges tend

[*] Corresponding Author. Email: ezgiyilmaz@hacettepe.edu.tr.

to assign exactly the same rating for each object (Tinsley and Weiss, 2005; Poppins, 2010). The agreement between objects rated independently by two raters or twice by the same rater is investigated with the agreement coefficients. There are different agreement coefficients for each differ- ent scale type (nominal, ordinal, and interval) and the type of coefficient changes according to the number raters. As a result, there is a huge literature on agree- ment coefficients. Although there are numerous agreement coefficients for each table structure or number of raters, there is no agreements on the use of these coefficients. Also, almost each coefficient has a specific disadvantage in calcula- tion or interpretation. In addition to agreement coefficients, approaches based on log-linear models for studying agreement patterns have also been proposed in the literature. There are specialized log-linear models for nominal and ordinal tables. Each model can be used to interpret the degree of agreement through odds ratios. In order to choose the most suitable way to evaluate agreement, we need to consider available literature on the context of agreement analysis and be aware of alternatives that provides accurate analysis of agreement.

In this article, considering the diversity of measures and approaches used to infer the degree and the direction of agreement and the importance of use of the most accurate tool for the evaluation of agreement, we present an extensive review of the literature on agreement coefficients and log-linear models used to evaluate agreement. We illustrate agreement coefficients calculated for two and multi- raters with nominal and ordinal categories, and also we mention disagreement coefficients. We present log-linear agreement models for nominal and ordinal cat- egories, and multi-rater studies. All of the discussed content are illustrated over three numerical examples.

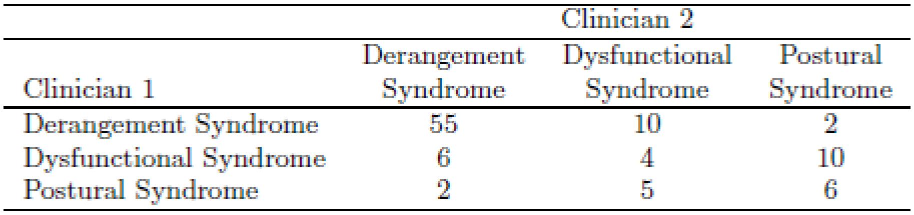
Two way and three way contingency table examples are cited in Section 2. The agreement coefficients are reviewed in Section 3. Section 4 presents the log-linear agreement models, followed by conclusion in Section 5.

2. Examples

In this section, we revisit three examples that will be used to illustrate the mea- sures and models related with the content of agreement.

Example 1: To illustrate the calculation of nominal agreement coefficients and agreement models, let us consider the square contingency table in Table 1. The data taken from Gwet (2012) who examined 100 individuals suffering from spinal pain. Two clinicians classified them in three categories according to their syn- drome type (e.g. Derangement, Dysfunction, or Postural). In this example, we investigate agreement between decisions of clinicians.

Table 1: Rating of spinal pain by Clinicians 1 and 2

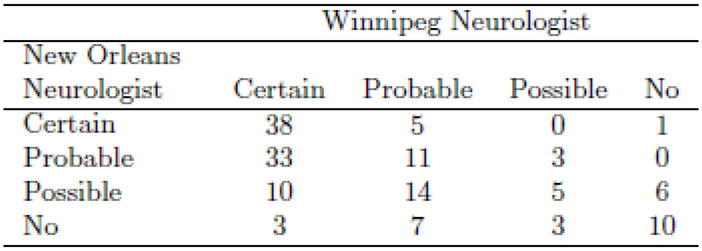


	Clinician 2		
Clinician 1	Derangement Syndrome	Dysfunctional Syndrome	Postural Syndrome
Derangement Syndrome	55	10	2
Dysfunctional Syndrome	6	4	10
Postural Syndrome	2	5	6

Example 2: 149 patients from Winnipeg are classified independently by two neurologists into four diagnostic categories in order to investigate the possibility that the disease was distributed differently geographically. The data is taken from Westlund and Kurland (1953) and also discussed by Landis and Koch (1977a), Gwet (2012), and Bangdiwala and Shankar (2013). To illustrate the calculation of ordinal agreement coefficients and agreement models, we consider the square contingency table in Table 2.

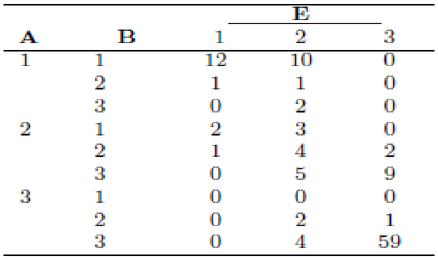
Example 3: The data in Table 3 is based on the data originally discussed by Holmquist, McMahon, and Williams (1967). This data set has also been analyzed in the studies of Landis and Koch (1977b), Becker and Agresti (1992), and Saracbasi (2011a). In order to investigate the variability in the classification of carcinoma in situ of the uterine cervix, three pathologists were classifying 118 slides into the 5 categories. Because the data contain sampling zero frequencies, the original categories are reclassified to the following categories: (1) Negative, (2) Atypical Squamous Hyperplasia, (3) Carcinoma in Situ + Squamous Carcinoma with Early Stromal Invasion + Invasive Carcinoma (Landis and Koch, 1977b; Becker and Agresti, 1992). This data set is used to illustrate the measures on models for multi-rater studies.

Table 2: Cross tabulations of multiple sclerosis diagnosis by two independent neurologists, comparing concordance with different sets of patients



	Winnipeg Neurologist			
New Orleans Neurologist	Certain	Probable	Possible	No
Certain	38	5	0	1
Probable	33	11	3	0
Possible	10	14	5	6
No	3	7	3	10

Table 3: Independent classification by three pathologists of most involved histological lesion



			E	
A	B	1	2	3
1	1	12	10	0
	2	1	1	0
	3	0	2	0
2	1	2	3	0
	2	1	4	2
	3	0	5	9
3	1	0	0	0
	2	0	2	1
	3	0	4	59

3. Agreement Coefficients

3.1 Agreement coefficients for nominal categories

The first approaches of agreement studies were focused on raw agreement which is equal to the observed proportion of agreement (von Eye, Schauerhuber, and Mair, 2007). Let n_{ij} denote the number of objects, n show the total number of observations, pi. indicate the ith row total probability, and $p_{.j}$ indicate the j^{th} column total probability in an R×R contingency table. Then the raw agreement, *ra* is calculated as the following where p_{ij} is the probability of cell (*i*, *j*) for *i*, *j* = 1, 2, ..., R:

$$ra = \sum_{i=1}^{R} P_{ii} \, . \tag{1}$$

Let P0 be the observed agreement equal to *ra* and *Pe* be the proportion agreement expected by chance, the general form for agreement coefficients is defined as the following (Zwick, 1988):

$$A = \frac{P_0 - P_e(A)}{1 - P_e(A)} \tag{2}$$

It is shown that, A coefficient given in Equation (2) provides a better description of the degree of agreement than *ra* (Zwick, 1988). Goodman and Kruskal (1954) suggested λ and Bennett, Alpert, and Goldstein (1954) suggested S coefficient. Bennett, Alpert, and Goldstein (1954) claimed that the proportion 1/R, where R is the number of categories, represents the best estimate of proportion agreement expected by chance (*Pe*) (Yang, 2007). Scott (1955) suggested π coefficient to overcome the defects of S. Scott (1955) argued that, ”It is convenient to assume that the distribution for the entire set of interviews represents the most probable (and hence ’true’ in the long-run probability sense) distribution for any individual coder.” Cohen (1960) discussed Scott’s π from the point that it ignores differences in rater marginals.

Cohen (1960) suggested κ statistics as a chance-corrected measure of agreement. The assumption of κ is that the ratings of raters are statistically independent and kappa allows different marginal probabilities of success associated with the raters to differ (Banerjee, Capozzoli ,McSweeney, Sinha, 1999). Cohen’s *K* coefficient is always applicable, easy to calculate and interpret, available in general purpose statistical software packages, and it condenses relevant information into one coefficient (Cohen, 1960). Oppositely, most authors discussed some limitations and insufficiencies of κ, 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 (Tanner and Young, 1985a; Kundel and Polansky, 2003). 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.

The coefficients S, π, and κ all have disadvantages. In formulating the chance corrections, the homogeneity of rater marginals is assumed by π and uniformity of marginals is represent by S (Zwick, 1988). Marginals are assumed to be fixed whenever the marginal probabilities are known to the rater before classifying the objects into categories. When the raters are completely free to assign objects to categories in any way they choose, the marginals are qualified ”free.” Brennan and Prediger (1981) claimed that κ is appropriate when marginal probabilities are fixed. If either or both of the marginals are free to vary, κ is replaced by S. Warrens (2010a) proved that S $\geq \pi \geq \lambda$ and κ $\geq \pi \geq \lambda$ for R × R tables and S is an upper bound of κ when matrix of the marginal probabilities are weakly symmetric.

Maxwell (1977) suggested the random error coefficient of agreement (RE) as a measure of agreement for 2 × 2 tables and Janes (1979) extended RE coefficient for R × R tables. When R > 2, the average disagreement (ad) is

$$ad = \frac{sum\ of\ proportions\ in\ disagreement\ cells}{R^2 - R}. \quad (3)$$

When the chance-corrected agreement for i^{th} category is $P_i = p_{ii} - ad$, RE is calculated as follows.

$$\text{RE} = P_0 + P_1 + \cdots + P_R. \quad (4)$$

Aickin (1990) suggested α coefficient and proposed an iterative algorithm to calculate the coefficient. Gwet (2008) suggested AC1 coefficient which is similar to κ in its formulation and its simplicity (in addition to being paradox-resistant) (Gwet, 2008; Gwet, 2012) and discussed the problem that the Pe of kappa differs from 0 to 1 despite the fact that Pe values should not exceed 0.5. Gwet discussed the necessity of a new formulation to compute the chance agreement probability (Gwet, 2012; Wongpakaran, Wongpakaran, Wedding, and Gwet, 2013). Wong- pakaran, Wongpakaran, Wedding, and Gwet (2013) concluded that in assessing the inter-rater reliability coefficient for personality disorders, Gwet’s AC1 is su- perior to Cohen’s κ and the results show that Gwet’s method over Cohen’s κ with regard to prevalence or marginal probability problem. Unlike κ and AC_1, the α coefficient is computation intensive. S has reappeared as the C coefficient of Janson and Vegelius (1979) and the $K\eta$ index of Brennan and Prediger (1981).

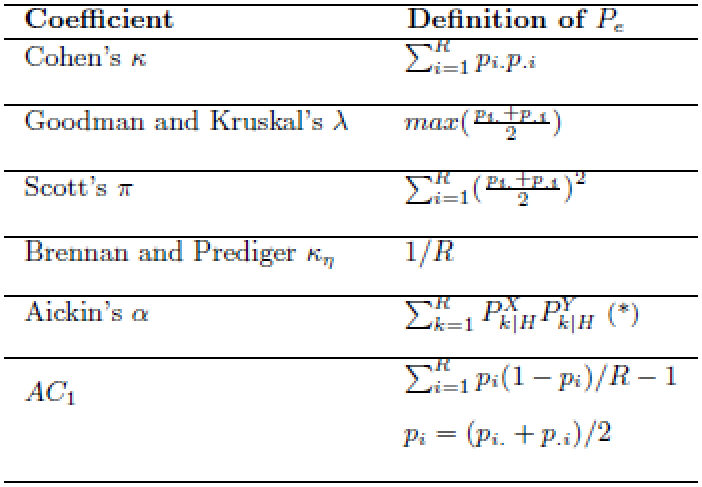
Although the coefficients are used to describe the agreement, they are based on different assumptions. Thus, they are not appropriate in all contexts. The assumptions are hidden in different definitions of *Pe* (Warrens, 2010a). These definitions are presented in Table 4 and the coefficients are calculated with Equation (2). For each of these coefficients, the formulation of *Pe* differs as seen in Table 4. 4. Here in the formulation of Aickin’s α, $P^X_{K|H}$ represents the probability for rater X to classify into category k, a subject known to be hard to classify (Gwet, 2012).

The Bangdiwala’s B_N statistic was derived from a graphical representation of R × R table, so it focuses on the area of agreement. It is calculated as

$$B_N = \frac{\sum_{i=1}^{R} n_{ii}^2}{\sum_{i=1}^{R} n_{i.}n_{.i}} \quad (5)$$

where $n_{i.}$ and $n_{.i}$ are the i^{th} row and j^{th} column totals, respectively. Because the BN statistic is a ratio, it ranges from 0 for no agreement to +1 for perfect agreement (Bangdiwala, 1988; Munõz and Bangdiwala, 1997)

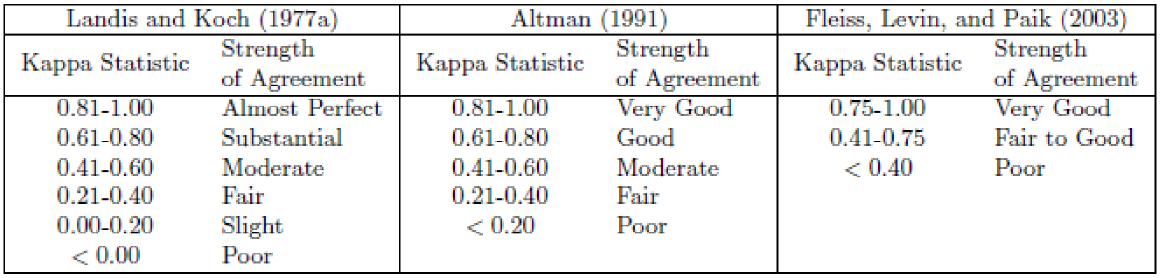
Table 4: Definitions of the proportion agreement expected by chance



Coefficient	**Definition of P_e**
Cohen's κ	$\sum_{i=1}^{R} p_{i.}p_{.i}$
Goodman and Kruskal's λ	$max(\frac{p_{i.}+p_{.i}}{2})$
Scott's π	$\sum_{i=1}^{R}(\frac{p_{i.}+p_{.i}}{2})^2$
Brennan and Prediger κ_η	$1/R$
Aickin's α	$\sum_{k=1}^{R} P_{k\|H}^{X} P_{k\|H}^{Y}$ (*)
AC_1	$\sum_{i=1}^{R} p_i(1-p_i)/R-1$ $p_i = (p_{i.}+p_{.i})/2$

In the literature, there are several interpretations of κ statistic. The inferences shown Table 5 can be assigned to the corresponding ranges of kappa (Landis and Koch, 1977a; Altman, 1991; Fleiss, Levin, and Paik, 2003).

Table 5: Interpretation of kappa statistics



Landis and Koch (1977a)		Altman (1991)		Fleiss, Levin, and Paik (2003)	
Kappa Statistic	Strength of Agreement	Kappa Statistic	Strength of Agreement	Kappa Statistic	Strength of Agreement
0.81-1.00	Almost Perfect	0.81-1.00	Very Good	0.75-1.00	Very Good
0.61-0.80	Substantial	0.61-0.80	Good	0.41-0.75	Fair to Good
0.41-0.60	Moderate	0.41-0.60	Moderate	< 0.40	Poor
0.21-0.40	Fair	0.21-0.40	Fair		
0.00-0.20	Slight	< 0.20	Poor		
< 0.00	Poor				

Table 6 shows the Munõz and Bangdiwala's (1997) summary of interpretation guidelines for κ and BN coefficients for 3 × 3 and 4 × 4 tables.

For Example 1, agreement coefficients are calculated and given in Table 7. The results show that the level of agreement is different for the coefficients. When kappa has the lowest agreement between clinicians decisions, *ra* has the highest level of agreement. As expected, Warrens's inequality ($K \geq \pi \geq \lambda$) is observed for this data. While it is possible to infer a fair agreement by *K*, it can be said that the agreement between clinicians is at a substantial on good level. This is a good example of discrepancy between measures and their interpretation.

Table 6: Interpretation of kappa and B_N statistics for 3 × 3 and 4 × 4 tables

P_0	Labels	Kappa*		B_N
1.0	Perfect	1.00		1.00
0.9	Almost Perfect	0.85	0.87	0.81
0.7	Substantial	0.55	0.60	0.49
0.5	Moderate	0.25	0.33	0.25
0.3	Fair	-0.05	0.07	0.09
0.1	Poor	-0.35	-0.20	0.01

*: In each row, the first value corresponds to 3 × 3 tables and the second value to 4 × 4 tables.

Table 7: The calculated agreement coefficients between clinicians

Coefficient	ra	κ	λ	π	κ_η	RE	AC_1	α	B_N
Estimate	0.650	0.322	0.000	0.321	0.475	0.475	0.528	0.401	0.636

For ordinal square tables, the hierarchy between levels of ordinal variables should be considered in the analysis of agreement. In that case, different coefficients focused in the next section should be used.

3.2 Agreement coefficients for ordinal categories

For ordinal categories, instead of kappa, weighted kappa coefficient is suggested for use (Cohen, 1968). The coefficient allows each (i, j) cell to be weighted ac- cording to the degree of agreement between ith and jth categories (Shoukri, 2004). Since the observed agreement and the proportion agreement expected by chance are

$$P_0 = \sum_{i=1}^{R}\sum_{j=1}^{R} \omega_{ij} p_{ij}, \tag{6}$$

and

$$P_0 = \sum_{i=1}^{R}\sum_{j=1}^{R} \omega_{ij} p_{i.} p_{.j}, \tag{7}$$

respectively, the weighted kappa coefficient $\widehat{K}_\omega$ is

$$K_\omega = \frac{P_0 - P_e}{1 - P_e}. \tag{8}$$

Vanbelle and Albert (2009) showed that using linear weights is equivalent to deriving a kappa coefficient from R − 1 embedded 2 × 2 tables. Bangdiwala (1988) also suggested the weighted version of B_N coefficient and suggested B_w coefficient. The sufficiency of weighted kappa was discussed by Warrens (2014). Warrens (2013a) discussed the kappa coefficients for 3 × 3 tables. Besides a variation of the weighted kappa, Kendall’s W coefficient was suggested to investigate interrater agreement (Kendall and Babington-Smith, 1939).

While all the disagreements accepted equal to calculate unweighted kappa coefficient, disagreements are ranked to calculate weighted kappa. The weights indicate disagreement and are used to calculate weighted kappa. Under this circumstance, selection of the weights has a great importance. Popular weights for weighted kappa are the linear and the quadratic weights shown in Equations (9) and (10), respectively (Cicchetti and Allison, 1971; Fleiss and Cohen, 1973). The quadratically and linearly weighted kappas are used for continuous-ordinal scale data. However, in practice, many scales are dichotomous ordinal. In this case, Warrens(2013b) suggested to use the additive weights shown in Equation (9). In recent studies, dispersion weights (Schuster and Smith, 2005), weights with the exponential and square distance functions shown in Equations (10) and (11) were suggested (Yang, 2007). It has been frequently observed in the literature that the value of the quadratically weighted kappa is higher than the value of the linearly weighted kappa (Warrens, 2012). This result implies that the level of the agreement depends on used weights. This is one of the disadvantages of weighted kappa.

- Linear weights:

$$\omega_{ij} = 1 - |i-j|/(R-1). \tag{9}$$

- Quadratic weights:

$$\omega_{ij} = 1 - (i-j)^2/(R-1)^2. \tag{10}$$

- Additive weights:

$$\omega_{ij} = \begin{cases} 0 & if\ i = j, \\ \sum_{l=i}^{j-1} \omega_l & i < j, \\ \sum_{l=j}^{i-1} \omega_l & i > j, \end{cases} \tag{11}$$

- Square distance function (SDF):

$$\omega_{\mathrm{ij}} = 1 - \frac{|i-j|^2}{R(R+1)^2(R+2)}\,. \tag{12}$$

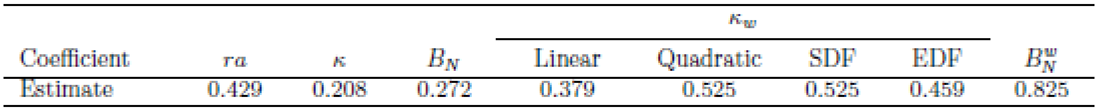
- Exponential distance function(EDF):

$$\omega_{\mathrm{ij}} = 1 - \frac{e^{|i-j|-1}}{\frac{e}{e-1}\left[\frac{e(e^R-1)}{1-R} + R\right] - \frac{R(R+1)}{2}}\,. \tag{13}$$

Berry and Mielke (1988) and Janson and Olsson (2001) discussed the agreement for the square tables with interval scale. Scott's π and Brennan and Prediger's k_n coefficients were generalized by Gwet (2012).

For Example 2, agreement coefficients for nominal and ordinal variables are calculated and given in Table 8. The results in Table 8 show that the coefficients for nominal variable are not consistent with those given for ordinal variables. Although the agreement is interpreted as slight from κ, it is obtained as moderate from *ra*. The level of agreement between neurologists for unweighted κ and $\mathrm{B_N}$ differ from weighted versions. Weighted kappa coefficient with linear weights has the lowest agreement between neurologists decisions and Bw has the highest agreement. The value of quadratically weighted kappa is found higher than linearly weighted kappa which has been remarked in literature.

Table 8: The calculated agreement coefficients between clinicians



				κ_w				
Coefficient	ra	κ	B_N	Linear	Quadratic	SDF	EDF	B_N^w
Estimate	0.429	0.208	0.272	0.379	0.525	0.525	0.459	0.825

3.3 Agreement coefficients for multi-rater studies

Cohen’s κ is suggested for use in two rater studies. For the multi-rater studies, Light’s κ (Light, 1971), which is the generalized form of Cohen’s κ, Hubert’s κ (1977), and Fleiss κ (1971) can be used (Shoukri, 2004; Warrens, 2010b). Hubert’s κ was independently proposed by Conger (1980). S coefficients were generalized for multi-raters by Randolph (2005). As an alternative to Fleiss κ, Gautam (2014) suggested A-kappa. Berry, Johnston, and Mielke (2008) suggested a kappa coefficient for ordinal square tables with multi-raters.

Let h be the number of raters, R be the number of categories, n be the number of observations, κ_{ij} in Equation (14) be the kappa coefficient among ith and j raters, and K_{ij} in Equations (15)-(17) be the number of raters that assign ith observation to category *j*. Then, Lights’s κ, Fleiss’s κ, Randolph’s S, and Gautam’s A-kappa coefficients are defined as follows:

- Light’s κ

$$\mathrm{L}(k) = \frac{2}{h(h-1)}\sum_{i=1}^{h-1}\sum_{i'=i+1}^{h} k_{ii'} \tag{14}$$

The measure L(κ) is the arithmetic mean of h(h − 1)/2 pairwise κ_{ii} that can be formed between *h* raters.

- Fleiss’s κ:

$$\mathrm{F}(\pi) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{R} k_{ij}^2 - hn\left[1+(h-1)\sum_{j=1}^{R} P_j^2\right]}{nh(h-1)\left[1-\sum_{j=1}^{R} P_j^2\right]}, \tag{15}$$

Where $\mathrm{P_j} = \frac{1}{hn}\sum_{i=1}^{n} K_{ij}$.

- Randolph's S:

$$\mathrm{R(S)} = \frac{\frac{1}{nh(h-1)}\left\{\sum_{i=1}^{n}\sum_{j=1}^{R} K_{ij}^2 - hn\right\} - \frac{1}{R}}{1-\frac{1}{R}}, \tag{16}$$

Where $i = 1,2,\dots,\mathrm{n}$ and $\mathrm{j} = 1,2\dots,\mathrm{R}$.

- Gautam's A-kappa (AK):

$$\overline{\mathrm{G}} = R\sum_{i=1}^{n}\sum_{j=1}^{R} K_{ij}^2/\left(nh^2(R-1)\right) - 1/(R-1) \tag{17}$$

$$\mathrm{AK} = \frac{\bar{G} - 1/h}{1 - 1/h} \tag{18}$$

Hubert's coefficient was rewritten for ordinal categories with different definitions of P_0^H and P_e^H (Warrens, 2010b). When p_i, q_j , and r_k are marginal proportions and A = $\{a_{ij}\}$, B = $\{b_{ij}\}$, and C = $\{c_{ij}\}$ are the sub-tables, given in Equation(19),

$$\alpha_{ij} = \sum_{k=1}^{R} P_{ijk} \quad b_{ij} = \sum_{k=1}^{R} P_{jik} \quad c_{ij} = \sum_{k=1}^{R} P_{jki} \;, \tag{19}$$

And

$$\mathrm{p}_i = \sum_{j=1}^{R}\sum_{k=1}^{R} P_{ijk} \quad q_i = \sum_{j=1}^{R}\sum_{k=1}^{R} P_{jik} \quad r_i = \sum_{j=1}^{R}\sum_{k=1}^{R} P_{jki} \;, \tag{20}$$

The P_0^H and P_e^H are defined as :

$$P_0^H = \frac{1}{3}\sum_{i=1}^{R}\sum_{j=1}^{R}\left[1 - \frac{|i-j|}{R-1}\right]\left(a_{ij} + b_{ij} + c_{ij}\right), \tag{21}$$

And

$$P_e^H = \frac{1}{3}\sum_{i=1}^{R}\sum_{j=1}^{R}\left[1 - \frac{|i-j|}{R-1}\right]\left(p_i q_j + p_i r_j + q_i r_j\right), \tag{22}$$

Berry, Johnston, and Mielke (2008) suggested a weighted kappa coefficient for ordinal 3-rater tables with the following P_0^M and P_e^M :

$$P_0^M = \sum_{i=1}^{R}\sum_{j=1}^{R}\sum_{k=1}^{R} \omega_{ijk} p_i q_j r_k \;, \tag{23}$$

And

$$P_e^M = \sum_{i=1}^{R}\sum_{j=1}^{R}\sum_{k=1}^{R} \omega_{ijk} p_i q_j r_k . \tag{24}$$

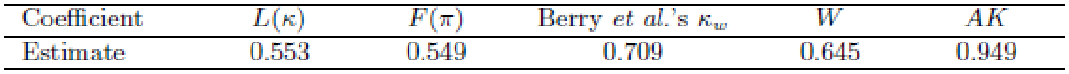
Here the weights w_{ijk} are calculated from Equation (25):

$$w_{ijk} = 1 - \frac{|i-j| + |i-k| + |j-k|}{2(R-1)} \;, \tag{25}$$

then the Hubert's and Berry's weighted kappas are calculated from Equation (2).

For Example 3, agreement coefficients for multi-rater tables are calculated and given in Table 9. The results show that the level of agreement between three pathologists is similar at L(κ) and F (π), and highest at Berry, Johnston, and Mielke's κw. Because the levels of carcinoma in situ of uterine cervix are ordinal, Berry, Johnston, and Mielke's κw is the most proper measure.

Table 9: The calculated agreement coefficients between pathologists



Coefficient	$L(\kappa)$	$F(\pi)$	Berry *et al.*'s κ_w	W	AK
Estimate	0.553	0.549	0.709	0.645	0.949

In addition to agreement coefficients, the coefficients that measure of disagreement mentioned in the next section are also suggested in the literature.

3.3 Disagreement coefficient

Cohen's k, Brennan and Prediger's k_η , and raw agreement coefficients were rewritten as disagreement measures (von Eye and von Eye, 2005). Since P d is the observed disagreement and P d is the proportion disagreement expected by chance, the raw disagreement is defined as the following:

$$\mathrm{ra}^d = 1 - P_0 = 1 - \sum_{i=j}^{R} Pii = P_0^d. \tag{26}$$

Cohen's k is rewritten as a disagreement measure (von Eye and von Eye, 2005):

$$P_e^d = \sum_{i\neq j}^{R} Pi.P.i = 1 - \sum_{i=j}^{R} Pi.P.i = 1 - Pe. \tag{27}$$

then the kappa is calculated from,

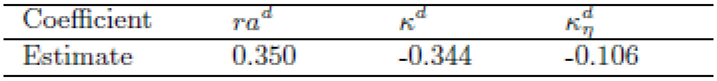
$$k^d = \frac{P_0^d - P_e^d}{1 - P_e^d} = \frac{-P_0 \;\; + P_e}{P_e} \tag{28}$$

Brennan ve Prediger's kη was rewritten as a disagreement measure:

$$k_n^d = \frac{\frac{1}{R} - P_0}{1/R}\,. \tag{29}$$

For Example 1, disagreement coefficients are calculated and given in Table 10. When the value of disagreement is positive, it can be said that there is a disagreement instead of agreement between raters. Here, because kd < 0 and kd < 0, instead of disagreement, there is more agreement between clinicians decisions. Because rad changes between 0 and 1, and the value of rad = 0.350, it can be said that there is more agreement than disagreement between the decisions of clinicians.

Table 10: The calculated disagreement coefficients between clinicians



Coefficient	ra^d	κ^d	κ_η^d
Estimate	0.350	-0.344	-0.106

4. Log-linear Agreement Models

Because of the insufficiency of agreement coefficients, most authors prefer to use log-linear agreement models. Instead of summarizing agreement, log-linear models analyze the structure of the agreement in the data (Tanner and Young, 1985a). Model studies give more detailed information about the table. In addition to analysis of agreement, odds ratios may be calculated under fitted model to infer the degree of agreement.

4.1 Agreement models for nominal categories

Agreement models are suggested to be used in square contingency tables with nominal categories. These are agreement (Tanner and Young, 1985a), disagreement, symmetric band disagreement (Tanner and Young, 1985b), and agreement plus disagreement (Saracbasi, 2011b) models. Consider an R × R contingency table that the first rater is represented by X and the second rater is represented by Y. In this two-way table, n subjects are cross-classified on two categorical responses. The corresponding log-linear model is as given in Equation (30).

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_l^Y + \delta_{ij}, \tag{30}$$

where λ is overall effect parameter, λ_i^X is the effect of X at *i* andλ_l^Y is the effect of Y at j with constraints such as $\sum_{i=1}^{R} \lambda_i^X = \sum_{j=1}^{C} \lambda_j^Y = 0$. m_{ij} 's are the expected values and δ_{ij} is the agreement parameter between X and Y, where i = 1, 2, . . . , R and j = 1, 2, . . . , C. The model is named with the agreement parameter. The agreement, disagreement, and symmetric band disagreement parameters are given in Equations (31), (32), and (33), respectively. The agreement and disagreement models have $(R - 1)^2 - 1$ degrees of freedom. The symmetric band disagreement model has $(R - 1)^2 - R + 1$.

$$\delta_{ij} = \begin{cases} \delta & if\ i = j, \\ 0 & otherwise. \end{cases} \tag{31}$$

$$\delta_{ij} = \begin{cases} \delta & if\ i \neq j, \\ 0 & otherwise. \end{cases} \tag{32}$$

$$\delta_{ij} = \begin{cases} \delta_1 & if\ |i-j| = 1, \\ \delta_2 & if\ |i-j| = 2, \\ & . \\ & . \\ & . \\ \delta_{R-1} & if\ |i-j| = R-1 \\ 0 & otherwise \end{cases} \tag{33}$$

The agreement plus disagreement model is,

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_l^Y + \lambda_{ij} + \delta_{ij}, \tag{34}$$

Where γ_{ij} is the agreement parameter that shown in the Equation (31) and δ_{ij} is the disagreement parameter that shown in the Equation (33).

Odds ratios (θ_{ij}) of successfully fitting models can be used to infer the agreement. The odds ratio diverges from 1 means that the decisions of raters are more similar than one level up decision, whereas the odds ratio converges to 0 means that the decisions of raters are more different than similar. The similarity indicates agreement between decisions of raters.

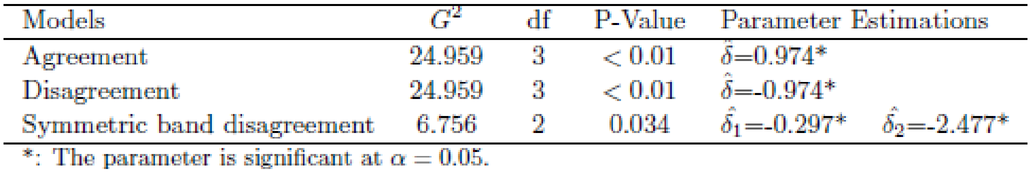
$$\theta_{ij} = \frac{m_{ij} m_{i+1,j+1}}{m_{i+1,j} m_{i,j+1}} \quad i = j = 1,2,\dots,R. \tag{35}$$

Agreement models are fitted to data in Example 1 and results are given in Table 11. While the calculated agreement coefficients in Table 7 indicate at least fair agreement between pathologists, estimated agreement coefficients (δ) also indicate the agreement. The agreement and disagreement models do not fit the data but symmetric band disagreement model fits the data at 1% level of significance. The agreement parameter in agreement model is $\delta > 0$ and found significant at 5% level of significance. Therefore, there is more agreement than expected by chance. Because $\delta < 0$ in disagreement models, there is less disagreement than expected by chance. In the symmetricband disagreement model, the disagreement parameters are both significant and this indicates agreement.

The best fitting model is found as symmetric band disagreement model. In this case, the odds ratios can be interpreted from the parameter estimates. The probability of giving derangement syndrome decision rather than dysfunctional syndrome decision (or giving dysfunctional syndrome decision rather than postural syndrome decision) of Clinician 1 are 1.81 times higher than derangement syndrome decision rather than dysfunctional syndrome decision (or giving dysfunctional syndrome decision rather than postural syndrome decision) of Clinician

2. The probability of giving derangement syndrome decision rather than dysfunctional syndrome decision (or giving dysfunctional syndrome decision rather than postural syndrome decision) of Clinician 1 are 6.57 times higher than dysfunctional syndrome decision rather than postural syndrome decision (or giving derangement syndrome decision rather than dysfunctional syndrome decision) of Clinician 2. Consequently, decisions of clinicians are more similar than one level up category and there is an agreement between them.

Table 11: The results of agreement models for the Example 1



Models	G^2	df	P-Value	Parameter Estimations	
Agreement	24.959	3	< 0.01	$\hat{\delta}$=0.974*	
Disagreement	24.959	3	< 0.01	$\hat{\delta}$=-0.974*	
Symmetric band disagreement	6.756	2	0.034	$\hat{\delta}_1$=-0.297*	$\hat{\delta}_2$=-2.477*

*: The parameter is significant at $\alpha = 0.05$.

4.2 Agreement models for ordinal categories

A way to apply agreement models for tables with ordinal variables is to ignore the hierarchy between adjacent categories of ordinal variables. However, this will lead loss of information. For an appropriate analysis of agreement for square contingency tables having ordered categories, association models with agreement parameter are suggested. In these models agreement and association are analyzed simultaneously.

The linear-by-linear association plus agreement model for two ordinal variables is:

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_l^Y + \beta u_i v_j + \delta_{ij}, \qquad (36)$$

Where $u_1 \leq u_2 \leq \ldots \leq u_R$ are ordered row scores and $v_1 \leq v_2 \leq \ldots \leq v_C$ are the column scores, and β is the association parameter. δij is the agreement parameter that is shown in the Equation (31). The linear-by-linear association plus agreement model has df = (R − 1)2 − 2. Goodman (1979) called the specifical case of uniform association plus agreement (UAA) model, where {u_i = i} and {v_j = j}. Bagheban and Zayeri (2010) called the model exponential scores association plus agreement, where $\{u_i = i^a\}$ and $\{vj = j^b\}$. Aktas and Saracbasi (2009) called the model symmetric disagreement plus uniform association (DUA), where the agreement parameter shown in Equation (37). This model has *df* = (R+1)(R−3).

$$\delta_{ij} = \begin{cases} \delta_1 & if\ |i-j| = 1, \\ \delta_2 & if\ |i-j| = 2, \\ \delta & if\ |i-j| \geq 3, \\ 0 & otherwise \end{cases} \tag{37}$$

In addition to uniform association, Valet, Guinot, and, Mary (2007) suggested non-uniform association plus agreement (NUAA) model to describe the variation of distinguishability between adjacent categories. Differently from the uniform association model, this model includes (R − 1) association parameters $\beta_{k,k+1}$ and extents the log-linear uniform association plus agreement model by allowing variations of distinguishability between adjacent categories. Thus, the model is useful to describe the quality of an ordinal scale more accurately. The model is:

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_l^Y - \frac{|i-j|}{2} \times \sum_{k=\min(i,j)}^{\max(i,j)-1} \beta_{k,k+1} + \delta_{ij}, \tag{38}$$

The agreement parameter is defined in the Equation (31). The model has df $=R^2 - 3R + 1$.

Fu, Gao, Tang, and Shi (2012) suggested a model combining ordinal scale information and category distinguishability between ordinal categories for modeling agreement. For this model, no score assignment is required for the ordinal categories.

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_l^Y + \lambda_{|i-j|}, \tag{39}$$

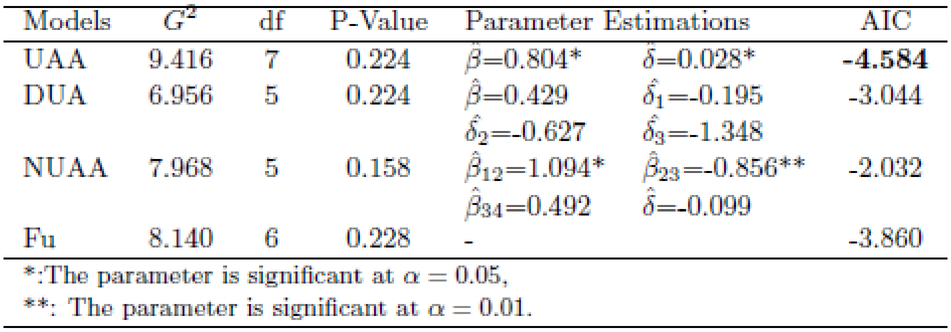
Where $0 = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{R-1}$. This model has *df* =(R-1)(R-2).

For Example 2, agreement models are calculated and given in Table 12. All the models fit the data. The association parameters are significant at 5% level of significance. Weighted kappas are at moderate level and B_N^ω is at almost perfect level, and the agreement parameters in the models are not significant.

Akaike Information Criteria (AIC = tt2 − 2*df*) is calculated for the models fit the data. The best fitting model is the model that has smallest AIC (Akaike, 1974). In that case, uniform association plus agreement (UAA) model is found as the best fitting model. According to the odds ratios from the matrix given in Equation (38), the odds ratios change depending on the distance to the main diagonal. The probability of giving same decision of New Orleans and Winnipeg Neurologists rather is 2.36 times higher than giving than one level up on decision. It means that, decisions of neurologists are more similar than one level up category. Thus, there is an agreement between them. The probability of giving certain decision rather than possible decision of New Orleans Neurologists is

2.17 times higher than probable decision rather than possible decision of Winnipeg Neurologist.

Table 12: Results of agreement models for the Example 2



Models	G^2	df	P-Value	Parameter Estimations		AIC
UAA	9.416	7	0.224	$\hat{\beta}$=0.804*	$\hat{\delta}$=0.028*	**-4.584**
DUA	6.956	5	0.224	$\hat{\beta}$=0.429	$\hat{\delta}_1$=-0.195	-3.044
				$\hat{\delta}_2$=-0.627	$\hat{\delta}_3$=-1.348	
NUAA	7.968	5	0.158	$\hat{\beta}_{12}$=1.094*	$\hat{\beta}_{23}$=-0.856**	-2.032
				$\hat{\beta}_{34}$=0.492	$\hat{\delta}$=-0.099	
Fu	8.140	6	0.228	-		-3.860

*:The parameter is significant at $\alpha = 0.05$,

**: The parameter is significant at $\alpha = 0.01$.

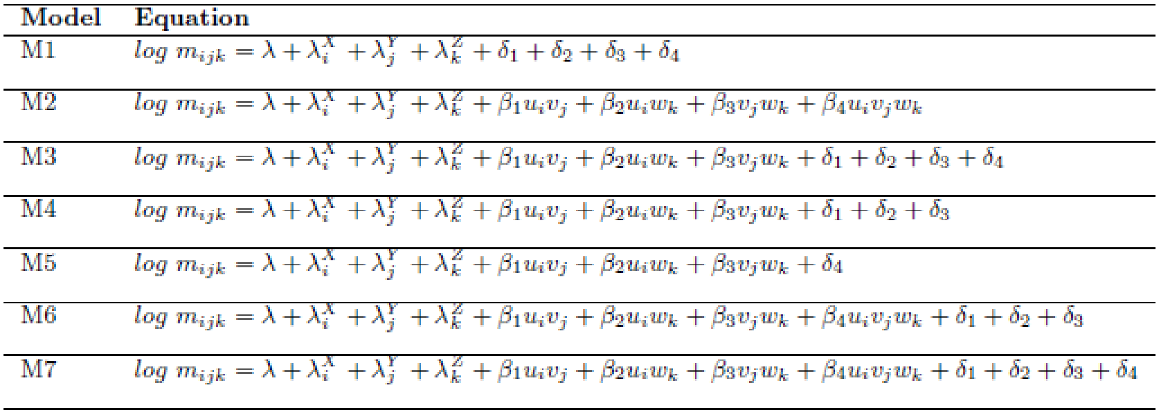
$$\hat{\theta}_{(UAA)} = \begin{bmatrix} 2.36 & 2.17 & 2.23 \\ 2.17 & 2.36 & 2.17 \\ 2.23 & 2.17 & 2.36 \end{bmatrix} \quad (40)$$

4.3 Agreement models for ordinal categories

For the multi-rater studies, global, global and partial, global and partial according to categories, and global and heterogeneous partial agreement models are suggested for nominal categories (Rogel, Boelle, and Mary, 1998; Kastango, 2006). Association plus agreement models are suggested for multi-rater studies with ordinal categories.

Let X, Y, and Z be the raters which have ordered categories, $u_i = i$, $v_j = j$, and $\omega_k = k$ are score values for variable X, Y, and Z, respectively. β_1 is the associa- tion parameter between X and Y, β_2 is between X and Z, β_3 is between Y and Z. δ is the global agreement parameter that shows the agreement between X, Y, and Z. For δ4, if i = j = k, δ_{ijk} is equal to 1. Association plus agreement models are shown in Table 13 where i = j = k = 1, 2, . . . , R (Melia and Diener-West, 1994; Lawal, 2003;, Saracbasi, 2011a).

Table 13: Uniform association plus agreement models models for multi-rater studies



Model	Equation
M1	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \delta_1 + \delta_2 + \delta_3 + \delta_4$
M2	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \beta_4 u_i v_j w_k$
M3	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \delta_1 + \delta_2 + \delta_3 + \delta_4$
M4	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \delta_1 + \delta_2 + \delta_3$
M5	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \delta_4$
M6	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \beta_4 u_i v_j w_k + \delta_1 + \delta_2 + \delta_3$
M7	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_1 u_i v_j + \beta_2 u_i w_k + \beta_3 v_j w_k + \beta_4 u_i v_j w_k + \delta_1 + \delta_2 + \delta_3 + \delta_4$

Let X, Y, and Z be the raters which have ordered categories for R × R × R contingency tables (R ≥ 3). As l = 1, 2, ..., (R − 1), $\beta_{l,l+1}$ is the association between the adjacent categories *l* and (*l + 1*) of X and Y, $\phi_{l,l+1}$ is the association between the adjacent categories *l* and (*l + 1*) of X and Z , $\omega_{l,l+1}$ is the association between the adjacent categories *l* and (*l + 1*) of Y and Z. Then, non-uniform association plus agreement models are shown in Table 14 (Yilmaz, 2013).

Table 15 shows goodness-of-fit test results of the models which were described in Table 13 and Table 14 for Example 3. Regarding the presented results, except M1 all models fit the data sufficiently well. The best fit belongs to the M2 that have the uniform association parameters between all pairs of pathologists and global association parameter and the second best fitting model is M10 that global association and agreement parameters between three pathologists.

Table 16 shows parameter estimates of M2 and M10 models. Although M2 is the best fitting model, the parameter estimates are not significant at 5% level of significance. The global agreement and association parameters of M10 are significant. The agreement parameter is $\delta < 0$ and significant, therefore, there is less agreement instead of expected by chance. Despite the agreement coefficients are at moderate level in Table 7, here the agreement parameter is negative which indicates disagreement.

Table 14: Non-uniform association plus agreement models for multi-rater studies

Model	Equation
M8	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{2} \times \sum_{l=min(i,j)}^{max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{2} \times \sum_{l=min(i,k)}^{max(i,k)-1} \varphi_{l,l+1}$ $- \frac{\lvert j-k\rvert}{2} \times \sum_{l=min(j,k)}^{max(j,k)-1} \omega_{l,l+1} - \frac{i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)}\beta$
M9	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{\lvert i-j\rvert}{2} \times \sum_{l=min(i,j)}^{max(i,j)-1} \beta_{l,l+1} - \frac{\lvert i-k\rvert}{2} \times \sum_{l=min(i,k)}^{max(i,k)-1} \varphi_{l,l+1}$ $- \frac{\lvert j-k\rvert}{2} \times \sum_{l=min(j,k)}^{max(j,k)-1} \omega_{l,l+1} - \frac{i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)}\beta + \delta_4$
M10	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)}\beta + \delta_4$
M11	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \frac{i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)}\beta + \delta_1 + \delta_2 + \delta_3$
M12	$log\ m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z - \lvert \frac{i-j\rvert+\lvert i-k\rvert+\lvert j-k\rvert}{2(R-1)}\beta + \delta_1 \rvert + \delta_2 + \delta_3 + \delta_4$

Table 15: Results of goodness-of-fit test for Example 3

Models	G^2	df	P-Value	AIC
M1	52.374	16	< 0.01	–
M2	7.222	16	0.969	**-24.778**
M3	5.983	13	0.947	-20.017
M4	5.983	14	0.967	-22.017
M5	9.581	16	0.888	-22.419
M6	3.453	13	0.996	-22.547
M7	3.453	12	0.991	-20.547
M8	7.106	13	0.897	-18.894
M9	3.306	10	0.973	-16.694
M10	12.870	18	0.799	-23.130
M11	9.882	16	0.873	-22.118
M12	8.478	15	0.903	-21.522

M2 and M10 models are found as the best fitting models. Because M10 has agreement parameter, odds ratios will be interpreted on this model parameters. The odds ratios for multi-way tables are called conditional odds ratios where one rater is accepted as fixed. For M10 model the conditional odds ratio matrices are $\hat{\theta}_{(i)jk} = \hat{\theta}_{i(j)k} = \hat{\theta}_{ij(k)}$ and given in Equation (37). The probability of giving atypical squamous hyperplasia decision rather than negative decision of pathologist B is 5.21 times higher than giving atypical squamous hyperplasia decision rather than negative decision of pathologist C for fixed levels of pathologist A. 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.

Table 16: The parameter estimates of M2 and M10

Models	Parameter Estimate	Standard Error	P-Value
M2	$\hat{\beta}_1 = 0.177$	0.842	0.834
	$\hat{\beta}_2 = 0.312$	0.962	0.745
	$\hat{\beta}_3 = 0.686$	0.867	0.428
	$\hat{\beta}_4 = 0.578$	0.371	0.119
M10	$\hat{\beta}_4 = 7.383$	1.514	0.000
	$\hat{\delta}_4 = -2.041$	0.863	0.018

$$\hat{\theta}_{(i)jk} = \hat{\theta}_{i(j)k} = \hat{\theta}_{ij(k)} = \left[\begin{array}{cc} 5.21 & 1.00 \\ 1.00 & 40.11 \\ \hdashline 5.21 & 7.70 \\ 7.70 & 5.21 \\ \hdashline 7.70 & 5.21 \\ 40.11 & 1.00 \end{array}\right] \tag{41}$$

5. Conclusion

In recent studies, interrater agreement analysis has grown extensively. There are different ideas between researchers when the subject is agreement. In practice, because coefficient of agreement summarize the rater agreement with a single number, some researchers prefer using coefficient of agreements, especially the kappa coefficient. Some researchers criticize the kappa coefficient in terms of loss information, undetermined weights, and undetermined interpretation. They assert to use agreement models instead of agreement coefficients. The main argument of the researchers who prefer to use agreement models reveals pure agreement. Odds ratios which are calculated from expected values of best fitting model are helpful to interpret the agreement in the square contingency tables.

In this paper, we present various methods for the study of interrater agreement when the response variable is nominal or ordinal categorical in the case that has two or multi raters. We focus on the agreement from the point of the coefficients and log-linear models.

We illustrate use of agreement coefficients and log-linear agreement models over nominal, ordinal, and multi-rater examples. The results show that all the agreement coefficients indicate different level of agreement and also log-linear model results differ. In that case, more than one coefficient depend on the scale of measurement should be considered and interpreted.

In fact, it would be appropriate to combine the results via meta analysis. Besides the agreement coefficients, agreement models should be applied to the data set. To draw more reliable inferences, κ coefficient which is calculated from the expected values of best fitting agreement model can be helpful to summarize the table with only one value

Acknowledgment

The authors would like to thank Dr. Haydar Demirhan for helpful comments and valuable suggestions on this article.

References

[1] Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen’s Kappa. Biometrics 46(2), 293-302..

[2] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716-723.

[3] Aktas, S. and Saracbasi, T. (2009). Estimation of symmetric disagreement using a uniform association model for ordinal agreement data. AStA 93(3), 335- 343..

[4] Altman, D. G. (1991). Practical statistics for medical research, 404. Chapman & Hall, London.

[5] Bagheban, A. A. and Zayeri, F. (2010). A Generalization of the uniform as- sociation model for assessing rater agreement in ordinal scales. Journal of Applied Statistics 37(8), 1265-1273.

[6] Bangdiwala, S. I. (1988). The agreement chart. Technical report, University of North Carolina at Chapel Hill, Department of Biostatistics, Institute of Statistics Mimeo, Series No. 1859 (Appendix).

[7] Bangdiwala, S. I. and Shankar, V. (2013). The agreement chart. BMC Medical Research Methodology 13(97), 1-7.

[8] Banerjee, M. et al. (1999). Beyond kappa: A review of interrater agreement measures. The Canadian Journal of Statistics 27(1), 3-23.

[9] Becker, M. P. and Agresti, A. (1992). A log-linear modelling of pairwise in- terobserver agreement on a categorical scale. Statistics in Medicine 11(1), 101-114.

[10] Bennett, E. M. et al. (1954). Communications through limited response ques- tioning. The Public Opinion Quarterly 18(3), 303-308.

[11] Berry, K. J. et al. (2008). Weighted kappa for multiple raters. Perceptual and Motor Skills 107(3), 837-848.

[12] Berry, K. J. and Mielke, P. W. (1988). A generalization of Cohen’s kappa agree- ment measure to interval measurement and multiple raters. Educational and Psychological Measures 48(4), 921-933.

[13] Brennan, R. L. and Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement 41, 687-699.

[14] Cicchetti, D. and Allison, T. (1971). A new procedure for assessing reliability of scoring eeg sleep recordings. American Journal EEG Technology 11(3), 101-109.

[15] Cicchetti, D. V. and Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology 43(6), 551-558.

[16] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37-46.

[17] Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70(4), 213-220.

[18] Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. Psychological Bulletin 88(2), 322-328.

[19] Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: The problems of the two paradoxes. Journal of Clinical Epidemiology 43(6), 543-549.

[20] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378-382.

[21] Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613-619.

[22] Fleiss, J. et al. (2003). Statistical methods for rates & proportions, 3rd edition, 598-626. Wiley & Sons, New York.

[23] Fu, L. et al. (2012). On modelling agreement and category distinguishability on an ordinal scale. Communications in Statistics-Theory and Methods 41, 4413-4426.

[24] Gautam, S. (2014). A-Kappa: A measure of agreement among multiple raters. Journal of Data Science 12, 697-716.

[25] Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. Journal of the American Statistical Association 49(268), 732-764.

[26] Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. Journal of the American Statisti- cal Association 74(367), 537-552.

[27] Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology 61, 29-48.

[28] Gwet, K. L. (2012). Handbook of inter-rater reliability, The definitive guide to measuring the extent of agreement among raters, 3rd edition. Advanced Analytics, LLC, Maryland.

[29] Holmquist, N. S. et al. (1967). Variability in classification of carcinoma in situ of the uterine cervix. Arch. Patholog. 84, 334-345.

[30] Hubert, L. (1977). Kappa revisited. Psychological Bulletin 84(2), 289-297.

[31] Janes, C. L. (1979). An extension of the random error coefficient of agreement to N × N tables. British Journal of Psychiatry 134, 617-619.

[32] Janson, H. and Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. Educational and Psychological Measurement 61(2), 277-289.

[33] Janson, S. and Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. Multivariate Behavioral Research 14(2), 255-269.

[34] Kastango, K. B. (2006). Assessing agreement among raters and identifying atypical raters using a log-linear modeling approach. Doctor of Philosophy, University of Pittsburg Graduate School of Public Health, Department of Biostatistics, Pittsburgh, 2006.

[35] Kendall, M. G. and Babington-Smith, B. (1939). The problem of m rankings. The Annals of Mathematical Statistics 10(3), 275- 287.

[36] Kundel, H. L. and Polansky, M. (2003). Measurement of observer agreement. Radiology 228(2), 303-308.

[37] Landis, J. R. and Koch, G. G. (1977a). The measurement of observed agreement for categorical data. Biometrics 33(1), 159-174.

[38] Landis, J. R. and Koch, G. G. (1977b). An application of hierarchical kappa- type statistics in the assessment of majority agreement among multiple observers. Biometrics 33(2), 363-374.

[39] Lawal, B. (2003). Categorical data analysis with SAS and SPSS applications (Edited by D. Riegert and J. Planer), 449-451, 490-492. New Jersey: Lawrence Erlbaum Associates, Publishers, Inc.

[40] Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin 76(5), 365-377.

[41] Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. British Journal of Psychiatry 130, 79-83.

[42] Melia, B. M. and Diener-West M. (1994). Modeling inter rater agreement for pathologic features of choroidal melanoma, in Case Studies in Biometry, (Edited by N. Lange et al.), Wiley and Sons.

[43] Muno˜z, S. R. and Bangdiwala, S. I. (1997). Interpretation of Kappa and B statistics measures of agreement. Journal of Applied Statistics 24(1), 105- 112.

[44] Poppins, R. (2010). Some views on agreement to be used in content analysis studies. Quality & Quantity 44, 1067-1078.

[45] Randolph, J. J. (2005). Free-marginal multirater kappa (multirater κfree): An alternative to Fleiss' fixed-Marginal multirater kappa.The Joensuu Learn- ing and Instruction Symposium, 14-15 October, Joensuu, Finland.

[46] Rogel, A. et al. (1998). Global and partial agreement among several observers. Statistics in Medicine 17, 489-501.

[47] Saracbasi, T. (2011a). Agreement models for multiraters. Turk J. Med. Sci. 41(5), 939-944.

[48] Saracbasi, T. (2011b). Agreement plus disagreement model for agreement data. Hacettepe Journal of Mathematics and Statistics 40(4), 609-616.

[49] Schuster, C. and Smith, D. A. (2005). Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. Psychome- trika 70(1), 135-146.

[50] Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. The Public Opinion Quarterly 19(3), 321-325.

[51] Shoukri, M. M. (2004). Measures of interrater agreement, 39-62. Florida: Chap- man & Hall/CRC Press LLC.

[52] Tanner, M. A. and Young M. A. (1985a). Modeling agreement among raters. JASA 80(389), 175-180.

[53] Tanner, M. A. and Young M. A. (1985b). Modeling ordinal scale disagreement. Psychological Bulletin 98(2), 408-415.

[54] Tinsley, H. E. A. and Weiss, D. J. (2000). Interrater reliability and agreement. Handbook of Applied Multivariate Statistics and Mathematical Modeling, (Edited by H. E. A. Tinsley and S. D. Brown), 94-124. New York: Academic Press.

[55] Valet, F. et al. (2007). Log-linear non-uniform association models for agreement between two ratings on an ordinal scale. Statistics in Medicine 26, 647-662.

[56] Vanbelle, S. and Albert, A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. Statistical Methodology 6, 157-163.

[57] von Eye, A. et al. (2007). Significance test for the measure of raw agreement. Interstat 1(1), 1-19.

[58] von Eye, A. and von Eye, M. (2005). Can one use Cohen's kappa to examine disagreement? Methodology 1(4), 129-142.

[59] Warrens, M. J. (2010a). Inequalities between kappa and kappa-like statistics for k × k tables. Psychometrica 75(1), 176-185.

[60] Warrens, M. J. (2010b). Inequalities between multi-rater kappas. Advanced Data Analysis Classification 4, 271-286.

[61] Warrens, M. J. (2012). Cohen's quadratically weighted kappa is higher than lin- early weighted kappa for tridiagonal agreement tables. Statistical Method- ology 9, 440-444.

[62] Warrens, M. J. (2013a). Weighted kappas for 3 ×3 tables. Journal of Probability and Statistics 1, 1-9.

[63] Warrens, M. J. (2013b). Cohen's weighted kappa with additive weights. Ad- vanced Data Analysis Classification 7, 41-55.

[64] Warrens, M. J. (2014). Power weighted versions of Bennett, Alpert and Gold- stein's S. Journal of Mathematics 2014, 1-9.

[65] Westlund, K. B. and Kurland, L. T. (1953). Studies on multiple sclerosis in Winnipeg, Manitoba, and New Orleans, Louisiana. II. A controlled investi- gation of factors in the life history of the Winnipeg patients. Am. J. Hyg. 57(3), 397-407.

[66] Wongpakaran, N. et al. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients:A study conducted with personality disorder samples. BMC Medical Research Methodology 13(61), 1-7.

[67] Yang, J. (2007). Measure of agreement for categorical data. Doctor of Philoso- phy, The Pennsylvania State University The Graduate School, Department of Statistics, PA.

[68] Yilmaz, A. E. (2013). Association models with agreement parameter for square contingency tables with ordered categories. Master of Science, Hacettepe University, Department of Statistics, Ankara.

[69] Zwick, R. (1988). Another look at interrater agreement. Psychological Bulletin, 103(3), 374-378, 1988.

Yilmaz Ayfer Ezgi
Department of Statistics
Hacettepe University
Ankara 06800, Turkey

Saracbasi Tulay
Department of Statistics
Hacettepe University
Ankara 06800, Turkey