# Multi-rater delta: extending the delta nominal measure of agreement between two raters to many raters

*Martín Andrés, A.*[1(1)] *and Álvarez Hernández, M.*[(2)]

(1) Bioestadística, Facultad de Medicina, Universidad de Granada. Granada, Spain.

(2) Centro Universitario de la Defensa - ENM, Universidad de Vigo, Vigo (Pontevedra), Spain.

## ABSTRACT

The need to measure the degree of agreement among *R* raters who independently classify *n* subjects within *K* nominal categories is frequent in many scientific areas. The most popular measures are Cohen's *kappa* (*R*=2), Fleiss' *kappa*, Conger's *kappa* and Hubert's *kappa* (*R*≥2) coefficients, which have several defects. In 2004, the *delta* coefficient was defined for the case of *R=2*, which did not have the defects of Cohen's *kappa* coefficient. This article extends the coefficient *delta* from *R*=2 raters to *R*≥2. The coefficient *multi-rater delta* has the same advantages as the coefficient *delta* with regard to the type *kappa* coefficients: *i)* it is intuitive and easy to interpret, because it refers to the proportion of replies that are concordant and non-random; *ii)* the summands which give its value allow the degree of agreement in each category to be measured accurately, with no need to be collapsed; and *iii)* it is not affected by the marginal imbalance.

**Keywords:** Cohen's *kappa*, Conger's *kappa*, *Delta* agreement, Fleiss' *kappa*, Hubert's *kappa*, nominal agreement.

## 1. Introduction

In many fields of science, including the behavioural sciences, geography and medicine, the degree of concordance or agreement among *R* raters that independently classify *n* subjects within *K* unordered categories is assessed (Fleiss, 1971; Landis & Koch, 1975a and b; Warrens, 2010; Schuster & Smith, 2005). When none of the raters is a gold-standard, the

[1] Correspondence to: Bioestadística. Facultad de Medicina. Universidad de Granada. 18071 Granada. Spain. E-mail: amartina@ugr.es. Phone:34-58-244080. Fax: 34-58-246117.

objective is to measure the degree of agreement between the two raters. When one of the raters is a gold-standard, the objective is to evaluate the degree of agreement of the problem rater with the gold-standard rater. In this article, we focus on the case in which no gold-standard rater exists.

Let us consider the case of two raters ($R=2$). Because some of the observed agreements may occur due to chance, the most common action is to eliminate the effect of chance using Cohen's *kappa* coefficient ($\kappa_C$) (1960). Although *Kappa* is a very popular and easily calculated measure of agreement, it has several disadvantages (Brennan & Prediger, 1981; Agresti *et al.*, 1995; Guggenmoos-Holzmann & Vonk, 1998; Nelson & Pepe, 2000; Martín Andrés & Femia Marzo, 2005; Erdmann *et al.*, 2015). The two most relevant disadvantages are its dependence on marginal distributions and the difficulty in measuring the degree of agreement for each category (even though the dependence on marginal distributions is seen as a desirable property by some authors: Vach 2005). Martín Andrés & Femia Marzo (2004, 2005 and 2008) proposed a response model that led to the measure of agreement *delta* ($\Delta$). Because it does not have the disadvantages of $\kappa_C$ (Ato *et al.*, 2011; Shankar and Bangdiwala 2014), it has been occasionally used in many different fields (Ecology, Geography, Psychology, Medicine, …). The *delta* model employs several measures that are valid in all circumstances, even if no gold-standard exists or if there exists, its marginal distribution is not fixed. However, the *delta* model is based on the assumption that one of the two raters is a gold-standard, which explains why its parameters are directly related to the situation. The model should be redefined for the case in which neither rater is a gold-standard, which ensures that the parameters of the model are directly related to the agreement parameters that are to be measured. This is the first aim of this article, which involves the *multi-rater delta* model and the agreement coefficient *delta* ($\Delta$). In addition, this extension will identify some minor errors committed by Martín Andrés & Femia Marzo when estimating the parameters of the *delta* model. A traditional example is Table 1(a), which comes from the classic example of Fleiss *et al.* (2003) where two raters diagnose 100 subjects in $K$=3 categories (Psychotic, Neurotic and Organic).

Now, let us consider the case where there are many raters ($R \geq 2$). Various authors have proposed generalizations of $\kappa_C$ for the multi-rater case (Fleiss, 1971; Hubert 1977; Conger 1980; Warrens, 2010; Marasini *et al.*, 2016). Although Fleiss' *kappa* ($\kappa_F$) is the most popular measure, it is not easy to interpret (Stoyan *et al.* 2012), nor is it an extension of $\kappa_C$ since the value of $\kappa_F$ is not equal to $\kappa_C$ when $R$=2 (Conger 1980); however, it is an extension of Scott's *pi* ($\pi$) coefficient (1955) (see the quotations from Warrens, 2010 and Marasini *et al.*, 2016) also called *intra-class kappa*. Conversely, the generalizations of Hubert's *kappa* (Hubert, 1977) and Conger's *kappa* (Conger 1980) are an extension of $\kappa_C$, particularly the two generalizations which will be noted later as $\kappa_{H2}$ (Hubert *pairwise*) and $\kappa_{HR}$ (Hubert *R-wise*). All multi-rater *kappa* coefficients exhibit a paradoxical behaviour (Marasini *et al.,* 2016; Conger, 2017), such as their dependence on the marginal imbalance. The second objective of this article is to extend the *delta* model from $R$=2 to $R \geq 2$, with the dual aim of obtaining a measure of agreement uninfluenced by the marginal distributions and which allows the degree of agreement in each class to be evaluated. An example for $R$=$K$=3 is listed in Table 2(a) from Dillon & Mulani (1984); it is cited by Schuster & Smith (2005) and "analyzed a persuasive communication study for which three raters classified 164 subjects as either positive, neutral, or negative"; however, in this paper, the ordinal quality will be analysed as if it were a nominal quality.

The analysis of the existing nominal agreement between several raters can be approached from two perspectives: (1) defining a statistic that measures the degree of agreement among the observers, which allows for a summary of measures of agreement (such as *kappa* statistics); and (2) defining a model for the agreements and disagreements observed, which allows more detailed analysis of the problem. The current *delta* model can be understood to be halfway between the two perspectives because, although a particular response model is adopted, the objective of the model is to measure the degree of overall agreement and the degrees of class-to-class agreement. Agresti (1992), Uebersax (1992), Banerjee *et al.* (1999) and Barlow W (2005) reviewed different agreement models. The options include the loglinear (Tanner and Young 1985a and b), association (Becker 1989, Goodman 1991) and quasi-symmetric models (Darroch and McCloud 1986, Agresti and Lang 1993), which model the expected frequencies. Another option is the Rasch model (Rasch 1961), which models the ratio of the classification probabilities into two classes of the same rater-subject pair. However, the most convenient option is the latent class model, which considers the existence of unobserved or latent variables (Dillon and Mulani 1984, Uebersax and Grove1990, Klauer and Batchelder 1996). The usual latent class model assumes that the rating level assigned by one rater is statistically independent of the rating levels assigned by other raters. This assumption of conditional independence can be unrealistic, so there are various methods to avoid it (Qu, Tan and Kutner 1996, Asselineau *et al.*, 2018). The current *delta* model is loosely related to the latent class model because its origin is in the model by Martín Andrés and Femia Marzo (2004, 2005), which in turn is derived from the model by Martín Andrés and Luna del Castillo (1989, 1990) for multiple choice tests (an extension of the classic model by Lord and Novick, 1968). And precisely the models of Martín Andrés and Luna del Castillo (1989) and the latent classes of Klauer & Batchelder (1996) are formally the same.

In the *delta* model for $R$=2, the parameter estimates and their standard errors (SEs) are obtained by the score method, which generates complicated deductions and expressions. Obtaining simpler equivalent expressions by a less cumbersome method, such as the classical multivariate *delta* method, is the third aim of this article. A free program (Multi-Rater Delta) may be downloaded to run the *multi-rater delta* model at http://www.ugr.es/local/bioest/software (Section "Agreement Among Raters").

This article is organized as follows: in section 2, the classic models *kappa* and *delta* are introduced, as well as the new *multi-rater delta* model for the case of $R$=2. In section 3, three classic *kappa* models are introduced as well as the new *multi-rater delta* model, for the $R\geq 2$ case. In section 4 the parameters of the *multi-rater delta* model and its standard errors (SE) are estimated. Section 5 offers three examples and finally, in section 6, the conclusions are set out.

## 2. Models for two raters

### *2.1. Kappa model*

Let us start with the case of two raters ($R=2$) that independently classify $n$ subjects within $K$ nominal categories. Given a subject, rater 1 classifies it as type $i$ ($i$ = 1, 2, ..., $K$) and rater 2 as type $j$ ($j$ = 1, 2, ..., $K$), which generates a table of absolute frequencies $x_{ij}$ and a table of relative frequencies (observed cell proportions) $\overline{p}_{ij}=x_{ij}/n$, with $\sum_{i=1}^{K}\sum_{j=1}^{K}x_{ij}=n$ and $\sum_{i=1}^{K}\sum_{j=1}^{K}\overline{p}_{ij}=1$. The notation for the observed totals of row ($x_{i\bullet}$ and $\overline{p}_{i\bullet}$), column ($x_{\bullet j}$ and $\overline{p}_{\bullet j}$) or overall ($x_{\bullet\bullet}=n$ and $\overline{p}_{\bullet\bullet}=1$) is typical; please refer to Table 1(a). If $p_{ij}$ is the probability that a subject is classified in cell ($i$, $j$), then the observed data set $\{x_{ij}\}$ is derived from a multinomial distribution of parameters $n$ and $\{p_{ij}\}$; $\{p_{i\bullet}\}$ and $\{p_{\bullet j}\}$ will be the marginal distributions of row raters and column raters, respectively.

To analyse the previous problem, Cohen (1960) defined the classic coefficient $\kappa_C=(I_o-I_e)/(1-I_e)$ estimated by $\hat{\kappa}_C=\left(\overline{I}_o-\overline{I}_e\right)/\left(1-\overline{I}_e\right)$, where $I_o=\sum_{i=1}^{K}p_{ii}$ ($\overline{I}_o=\sum_{i=1}^{K}\overline{p}_{ii}$) is the real (estimated) observed agreement index, and $I_e=\sum_{i=1}^{K}p_{i\bullet}p_{\bullet i}$ ($\overline{I}_e=\sum_{i=1}^{K}\overline{p}_{i\bullet}\overline{p}_{\bullet i}$) is the real (estimated) expected agreement index; the estimated values have been obtained on the assumption of independence between the classifications of the two raters. The denominator $\left(1-\overline{I}_e\right)$ has the two following effects on $\hat{\kappa}_C$ (Martín Andrés & Femia Marzo 2004); similarly with $(1-I_e)$ for $\kappa_C$. On the one hand its value is very dependent on the marginal distributions. On the other, its interpretation is not absolute, but relative to the value of $\left(1-\overline{I}_e\right)$; for the example in Table 1(a) one obtains $\hat{\kappa}_C$=.6765, indicating that "the two raters agree on 67.65% of the maximum number possible of non-random agreements".

*2.2. Delta model (classic)*

One way to correct these two defects is by using the following *delta* model (Martín Andrés & Femia Marzo, 2004):

$$p_{ij}/p_{i\bullet} = \delta_{ij}\Delta_i+(1-\Delta_i)\pi_j\ , \qquad \textbf{(1)}$$

where $\delta_{ij}$ allude to the Kronecker delta, $0\leq\pi_j\leq1$, $\sum_{j=1}^{K}\pi_j=1$, $\Delta_i\leq1$ and $\Delta=\sum_{i=1}^{K}\Delta_i\leq1$. This model is based on the model Martín Andrés and Luna del Castillo (1989) used in the context of multiple choice tests –see their expression (3)– which in turn is based on the classic model by Lord and Novick (1968); the model of Martín Andrés and Luna del Castillo (1989) is the same as the model of the expression (1) of Klauer and Batchelder (1996), regardless of the name given to each parameter. This is why the response model by the current expression (1) can be interpreted as follows. When rater 2 faces a subject classified as type *i* by rater 1 (who is assumed to be a gold-standard rater), s/he recognizes the subject with probability $\Delta_i$; when it is recognized, s/he correctly classifies it. When s/he does not recognize it, s/he does so with a probability $1-\Delta_i$, randomly classifies it as type *j* with a probability $\pi_j$. Given that $p_{ii}=p_{i\bullet}\Delta_i+p_{i\bullet}(1-\Delta_i)\pi_i$, the proportion of agreements in category *i* is the sum of the proportions of agreements that are not due to chance -the first summand $p_{i\bullet}\Delta_i$- and the proportions of agreements due to chance -the second summand $p_{i\bullet}(1-\Delta_i)\pi_i$. Hence, the total proportion of agreements not occurring by chance will be $\Delta=\Sigma p_{i\bullet}\Delta_i$. The authors (Martín Andrés & Femia

Marzo 2005) found that the value of the estimator $\hat{\Delta}$ of $\Delta$ is usually very similar to the value of $\hat{\kappa}_C$, except in certain cases in which at least one marginal distribution is very unbalanced. In the example in Table 1(a) one obtains $\hat{\Delta}$=.6875, a very similar value to that of $\hat{\kappa}_C$ despite the imbalance of the two marginals, but in other cases the values can be very different. However, the interpretation is now more direct: 68.75% of the responses are concordant beyond chance (as opposed to 89%=75%+4%+10% of raw agreements).

*2.3. Delta model (new)*

It can be seen that the classic *delta* model is formulated under the idea that the row rater is a gold-standard, even when the model is also otherwise valid. To eliminate this dependence, in Appendix A it is proved that the classic *delta* model is equivalent to the following new *delta* model:

$$p_{ij} = \delta_{ij}\alpha_i + (1-\Delta)\pi_{i1}\pi_{j2} \quad \mathbf{(2)}$$

where $\alpha_i \leq 1$, $\Delta = \Sigma\alpha_i \leq 1$ (which is the same parameter $\Delta$ as in the classic *delta* model), $0 \leq \pi_{i1} \leq 1$, $0 \leq \pi_{j2} \leq 1$ and $\sum_{i=1}^{K}\pi_{i1} = \sum_{j=1}^{K}\pi_{j2} = 1$. The following response model is assumed in the new *delta* model (see Appendix A), a very different model from the latent class model by Klauer and Batchelder (1996). When the two raters are faced with a given subject, both raters recognize it as category *i* with a probability of $\alpha_i$; when they do recognize it, they classify it as type *i*; when they do not recognize it, they do so with probability $1-\Delta$, classifying it randomly and independently with probability distributions $\{\pi_{i1}\}$ and $\{\pi_{j2}\}$, respectively.

If no model is considered, there exist three "raw" parameters of interest. The first two parameters are the proportion of agreements $p_{ii}$ in category *i* and the total proportion of agreements $p = \sum_{i=1}^{K} p_{ii}$. The parameter *p* measures the degree of total agreement; however, the parameter $p_{ii}$ does not measure the degree of agreement in category *i* but does measure the contribution of category *i* to the total agreement, since $p_{ii}$ is dependent on the degree of agreement in category *i* and the marginal distributions $p_{i\bullet}$ and $p_{\bullet i}$ in category *i*. A suitable method for defining the degree of agreement in category *i* (the third parameter) is through the *consistency* $S_i=2p_{ii}/(p_{i\bullet}+p_{\bullet i})$ of Martín Andrés & Femia Marzo (2005), which is a parameter that is based on the "proportion of specific agreement" of Fleiss *et al.* (2003) and on the agreement index of Cichetti & Feinstein (1990). The parameter $S_i$ measures the proportion of agreements among all subjects classified in category *i* by either of the two raters. The three agreement parameters $p_{ii}$, $p$ and $S_i$ are raw parameters since they are defined without considering the effect of chance. Since $p_{ii}=\alpha_i+(1-\Delta)\pi_{i1}\pi_{i2}$ from expression (2), then $p_{ii}$ is the sum of the proportion of agreements $\alpha_i$ that do not occur by chance and the proportion of agreements $(1-\Delta)\pi_{i1}\pi_{i2}$ that do occur by chance. If in the three previous raw parameters $p_{ii}$ are replaced by $\alpha_i$, then three parameters that are corrected for chance will be obtained (which is the current objective). In the new *delta* model, the consequence is that the three parameters of interest are $\alpha_i$ (the proportion of agreement in category *I* that do not occur by chance), $\Delta$ (the total proportion of agreements that do not occur by chance, that is, the overall degree of agreement), and *consistency* $\mathcal{S}_i$ in category *i* (the degree of agreement in category *i* that do not occur by chance),

$$\mathcal{S}_i=\frac{2\alpha_i}{p_{i\bullet}+p_{\bullet i}}=\frac{2\alpha_i}{2\alpha_i+(1-\Delta)(\pi_{i1}+\pi_{i2})}, \qquad \textbf{(3)}$$

where the second equality is due to $p_{i\bullet}=\sum_{j=1}^{K}p_{ij}=\alpha_i+(1-\Delta)\pi_{i1}$ and $p_{\bullet j}=\sum_{i=1}^{K}p_{ij}=\alpha_j+(1-\Delta)\pi_{j2}$. Parameter $\alpha_i$ (the proportion of agreements in category *i* that are not random) is not of primary interest. In the new *delta* model, note that the possibility that $\mathcal{S}_i<0$, $\alpha_i<0$ or $\Delta<0$ is allowed because the agreement can sometimes be negative. This can be interpreted as one of the raters classifying subjects "the other way around" compared to the other rater; that is, if the subject is in category *i* for one of the raters, the subject is in a category other than *i* for the other rater. Martín Andrés & Femia Marzo (2005) explained that $\mathcal{S}_i$ has the same objective as that pursued when defining $\kappa_C$ for collapsed data in category *i* (parameter $\kappa_{C(i)}$), which is an aim that is always achieved with $\mathcal{S}_i$; however, sometimes it is not achieved with $\kappa_{C(i)}$. In fact, a total agreement parameter based on the collapsed data in category *i* measures the degree of total agreement in the new situation, but does not measure the degree of agreement in category *i* because it should also measure the degree of agreement in the "not *i*" category.

## 3. Models for many raters

### *3.1. Multi-rater kappa*

Let there now be $R\geq 2$ raters who independently classify *n* subjects in *K* categories, producing a data matrix $\{y_{sr}\}$, with $s=1, 2, \ldots, n$, $r=1, 2, \ldots, R$ and $y_{sr}=1, 2, \ldots, K$; in this matrix $y_{sr}=i$ when the rater *r* classifies subject *s* into category *i*. The most usual thing to do is to summarize this information in a table of absolute frequencies $x_{i_1 i_2 \ldots i_R}=\#\{s \mid y_{s1}=i_1, \ldots, y_{sR}=i_R\}$ of dimension $K^R$, where the symbol # refers to "cardinal" and $x_{i_1 i_2 \ldots i_R}$ is the number of subjects classified as type $i_1$ by rater 1, type $i_2$ by rater 2, ..., or type $i_R$ by rater *R*. With this classification, $\{x_{i_1 i_2 \ldots i_R}\}$ are the observed values of a multinomial random variable of sample size *n* and probabilities $\{p_{i_1 i_2 \ldots i_R}\}$. Let $\overline{p}_{i_1 i_2 \ldots i_R} = x_{i_1 i_2 \ldots i_R}/n$ be the observed cell proportion, $\overline{p}_i = \overline{p}_{ii\ldots i}=\#\{s \mid y_{s1}=\ldots= y_{sR}=i\}/n=x_{ii\ldots i}/n=x_i/n$ the observed proportion of agreements in category *i* and its average value $p_i=p_{ii\ldots ii}$, $\overline{p}=\sum_{i=1}^{K}\overline{p}_i$ the observed total proportion of agreements and its average value $p=\sum_{i=1}^{K}p_i$, and finally $\overline{t}_{ir}=\sum_{i_1=1}^{K}\cdots\sum_{i_{r-1}=1}^{K}\sum_{i_{r+1}=1}^{K}\cdots\sum_{i_R=1}^{K}\overline{p}_{i_1\ldots i_r=i\ldots i_R}=\#\{s \mid y_{sr}=i\}/n$ the observed total proportion of responses *i* of rater *r* and its average value $t_{ir}=\sum_{i_1=1}^{K}\cdots\sum_{i_{r-1}=1}^{K}\sum_{i_{r+1}=1}^{K}\cdots\sum_{i_R=1}^{K}p_{i_1\ldots i_r=i\ldots i_R}$.

The $\kappa_C$ coefficient can be generalized to the case of multi-rater in several ways, depending on how the phrase “an agreement occurs” is interpreted. If $A_0$ is the observed number of agreements, Max $A_0$ is the maximum number and E($A_0$) is the average value of $A_0$ on the assumption of independence among all the raters, then Hubert (1977) indicated that the estimate of the degree of agreement of type *kappa* is given by

$$\hat{\kappa} = \frac{A_0 - \mathrm{E}(A_0)}{\mathrm{Max}(A_0) - \mathrm{E}(A_0)} = 1 - \frac{\mathrm{Max}(A_0) - A_0}{\mathrm{Max}(A_0) - \mathrm{E}(A_0)} = 1 - \frac{1-\overline{I}_o}{1-\overline{I}_e} \qquad \textbf{(4)}$$

where $\overline{I}_o$=$A_0$/Max($A_o$) and $\overline{I}_e$=E($A_0$)/Max($A_o$). Hubert (1977) makes the following interpretation "an agreement occurs if and only if all raters agree on the categorization of an object" or DeMoivre’s definition of agreement. In this case, Max ($A_o$)=$n$, $A_o$= $n\overline{I}_o = n\sum_{i=1}^{K}\overline{p}_i = n\overline{p}$ , E($A_o$)=$n\overline{I}_e$=$n\sum_{i=1}^{K}\prod_{r=1}^{R}\overline{t}_{ir}$ and Hubert's *kappa* coefficients is

$$\hat{\kappa}_{HR} = \frac{\overline{I}_o - \overline{I}_e}{1-\overline{I}_e},\ \text{where}\ \overline{I}_o = \sum_{i=1}^{K}\overline{p}_i = \overline{p}\ \text{and}\ \overline{I}_e = \sum_{i=1}^{K}\Pi_{r=1}^{R}\overline{t}_{ir}, \qquad \textbf{(5)}$$

which is an estimator of the population coefficient $\kappa_{HR}$=($I_o$–$I_e$)/(1–$I_e$), where $I_o = \sum_{i=1}^{K} p_i = p$ and $I_e = \sum_{i=1}^{K}\Pi_{r=1}^{R}t_{ir}$. Martín Andrés and Álvarez Hernández (2019) recently analysed this measure and obtained its SE.

However, the most traditional approach to understanding the phrase "an agreement occurs" is to understand the phrase "an agreement occurs if, and only if, two raters categorize an object consistently" by Fleiss (1971) and Hubert (1977) or a pairwise definition of agreement. An extension of the concept is Conger's $g$-wise *kappa* (1980), with 2≤$g$≤$R$, where $g$=$R$ or $g$=2 yields the two previously mentioned Hubert definitions (DeMoivre or $R$-wise and pairwise or pairwise, respectively). However, *kappa* coefficients can vary from one author to another, depending on the definition of $I_e$. The most traditional definitions are those of the Fleiss *kappa* ($\kappa_F$) (Fleiss 1971) and that of Hubert’s *kappa* ($\kappa_{H2}$) estimated by

$$\hat{\kappa}_F = 1 - \frac{nR^2 - \sum_{s=1}^{n}\sum_{i=1}^{K}R_{si}^2}{nR(R-1)\left\{1-\sum_{i=1}^{K}R_i^2\right\}}\ \text{and}\ \hat{\kappa}_{H2} = 1 - \frac{R^2 - \left(\sum_{s=1}^{n}\sum_{i=1}^{K}R_{si}^2/n\right)}{R(R-1) - 2\sum_{i=1}^{K}\sum_{r=1}^{R}\sum_{r'=r+1}^{R}\overline{t}_{ir}\overline{t}_{ir'}}, \qquad \textbf{(6)}$$

where $R_{si}$=#$\{r \mid y_{sr}=i\}$=0, 1, …, $R$ is the number of raters that classify subject $s$ in category $i$ and $R_i = R_{\bullet i}/nR$ is the total proportion of responses $i$ (any rater). In the case of $\hat{\kappa}_{H2}$, the expression derives from the fact that $A_o$=$\sum_{s=1}^{n}\sum_{i=1}^{K} R_{si}\left(R_{si}-1\right)/2$, Max($A_0$)=$nR(R-1)/2$ and E($A_o$)=$n\sum_{i=1}^{K}\sum_{r=1}^{R}\sum_{r'=r+1}^{R}\overline{t}_{ir}\overline{t}_{ir'}$. Warrens (2010) proved that $\hat{\kappa}_F \le \hat{\kappa}_{H2}$ and that if all the raters classify all the subjects $\hat{\kappa}_{H2}$ is more appropriate. It can be seen that when $R$=2 then $\hat{\kappa}_{HR} = \hat{\kappa}_{H2} = \hat{\kappa}_C$ but $\hat{\kappa}_F \ne \hat{\kappa}_C$.

*3.2. Multi-rater delta*

The extension of the *multi-rater delta* model for $R$=2 -expressions (2)- to the case of many raters is immediate. Now

$$p_{i_1 i_2 \ldots i_R} = \delta_{i_1 i_2 \ldots i_R}\alpha_{i_1} + \left(1-\Delta\right)\prod_{r=1}^{r=R}\pi_{i_r r} \text{ with } \Delta = \sum_{i=1}^{K}\alpha_i \qquad \textbf{(7)}$$

where $i_r$=1, 2, …, $K$, $\Delta$≤1 and $\alpha_i$≤1 and 0≤$\pi_{i_r r}$≤1 are the parameters of the *multi-rater delta* model. The parameters of interest, with similar interpretations to the interpretations of case $R=2$, are $\alpha_i$, $\Delta$ and the following extension of expression (3)

$$\mathcal{S}_i = \frac{R\alpha_i}{p_{i_1\bullet\ldots\bullet}+\ldots+p_{\bullet\ldots\bullet i_R}} = \frac{R\alpha_i}{R\alpha_i + \left(1-\Delta\right)\sum_{r=1}^{R}\pi_{ir}} \qquad \textbf{(8)}$$

where $p_{i_1\bullet\ldots\bullet} = \sum_{i_2=1}^{K}\ldots\sum_{i_R=1}^{K} p_{i_1 i_2 \ldots i_R}$ etc. In the *multi-rater delta* model, note that the possibility that $\mathcal{S}_i$<0, $\alpha_i$<0 or $\Delta$<0 is also allowed because the degree of agreement can sometimes be negative. Again, $\alpha_i$ is not a parameter of primary interest.

It can be seen that the coefficient $\Delta$ can be put into the traditional *kappa* format; since $p_i$=$\alpha_i$+(1−$\Delta$)$\sum_{r=1}^{R}\pi_{ir}$ through expression (7), then by adding up in $i$ and working out $\Delta$ we obtain

$$\Delta = \frac{I_o - I_\pi}{1 - I_\pi} \text{ where } I_\pi = \textstyle\sum_{i=1}^{K}\prod_{r=1}^{R}\pi_{ir}, \qquad \textbf{(9)}$$

so that the expected index of agreements under the *multi-rater delta* model is obtained on the basis of the probabilities $\pi_{ir}$ instead of on the basis of the probabilities $t_{ir}$ of the marginal distributions.

## 4. Estimation with the *delta* model

### *4.1. General case of more than two raters or more than two categories (R*>2 or *K*>2)

To make inferences with the *multi-rater delta* model, only some of the observed cell proportions $\overline{p}_{i_1 i_2 \ldots i_R}$ are needed: $\overline{p}_i$ (the observed proportions of agreements in category *i*) and $\overline{d}_{ir} = \overline{t}_{ir} - \overline{p}_i$ (the observed proportions of disagreements by rater *r* in category *i*). Based on these proportions, the total proportion of agreements $\overline{p} = \sum_{i=1}^{K} \overline{p}_i$, the total proportion of disagreements $\overline{D} = \sum_{i=1}^{K} \overline{d}_{ir}$ (which is the same for all raters), and the total proportion of disagreements in category *i* $\overline{D}_i = \sum_{r=1}^{R} \overline{d}_{ir}$ are determined. Based on the definitions, $\overline{t}_{ir} = \overline{p}_i + \overline{d}_{ir}$, $1 = \overline{p} + \overline{D}$, $\overline{D} = \sum_{i=1}^{K} \overline{D}_i / R$ and $\overline{D}_i \le (R-1)\overline{D}$. In addition, it can be seen that the values $R_i$ of $\hat{\kappa}_F$ are given by $R_i = \overline{p}_i + \overline{D}_i / R$.

Once these proportions are known, Appendix B shows that the estimators of the maximum likelihood of the various parameters are as follows:

$$\hat{\pi}_{ir} = \frac{\lambda_i + \overline{d}_{ir}}{B},\ \hat{\alpha}_i = \overline{p}_i - \lambda_i,\ \hat{\Delta} = 1 - B \text{ and } \hat{\mathcal{S}}_i = \frac{R\hat{\alpha}_i}{\overline{N}_i} = \frac{R\hat{\alpha}_i}{R\hat{\alpha}_i + \left(1-\hat{\Delta}\right)\sum_{r=1}^{R} \hat{\pi}_{ir}} \qquad \textbf{(10)}$$

with $\overline{N}_i = R\overline{p}_i + \overline{D}_i$, where $B \geq 0$ and $\lambda_i \geq 0$ are the solutions of expressions

$$B^{R-1} = \frac{\prod_{r=1}^{R}\left(\lambda_i + \overline{d}_{ir}\right)}{\lambda_i} \ (\forall i \mid \overline{d}_{ir} \neq 0,\ \forall r) \text{ under the condition } g(B) = \sum_{i=1}^{K} \lambda_i - B + \overline{D} = 0 \qquad \textbf{(11)}$$

with the exception that $\lambda_i = 0$ when $\overline{d}_{ir} = 0$ for some rater *r*. In addition, the *Supplementary Material* explicitly sets out how to proceed in order to determine the value of *B*, with details on how to act in the extreme cases of $B=0$ and $B=\infty$. In the particular case of independence among raters, that is, $\overline{p}_i = \prod_{r=1}^{R} \overline{t}_{ir} = \prod_{r=1}^{R}\left(\overline{p}_i + \overline{d}_{ir}\right)$, then $\lambda_i = \overline{p}_i$ and $B$=1 are the solutions of expressions (11); thus, $\hat{\alpha}_i = \hat{\Delta} = \hat{\kappa}_{HR}$=0. When $\overline{I}_o = \sum_{i=1}^{K} \overline{p}_i$=1, in which case $\hat{\kappa}_{HR}$=1, then $\overline{d}_{ir}$=0 ($\forall i$, $r$), $\lambda_i$=0 ($\forall i$), $g(B)$=0 implies that $B=\overline{D}$=0 and $\hat{\alpha}_i = \hat{\Delta} = \hat{\kappa}_{HR}$=1. In general, $\hat{\Delta}$ and $\hat{\kappa}_{HR}$ take similar values, except when the marginals are very unbalanced, in which case $\hat{\Delta}$ is usually superior to $\hat{\kappa}_{HR}$ (please see the examples in section 5 and at the end of Appendix B).

Once the parameter estimates of the *multi-rater delta* model are known, the classical chi-square test of the goodness of fit of the model can be performed. Since the observed quantities and expected quantities are $n\overline{p}_{i_1 i_2 \ldots i_R}$ and $n\hat{p}_{i_1 i_2 \ldots i_R} = n\delta_{i_1 i_2 \ldots i_R}\hat{\alpha}_{i_1} + n\left(1-\hat{\Delta}\right)\prod_{r=1}^{r=R}\hat{\pi}_{i_r r}$, respectively, then the type I $\alpha$ error test consists of comparing

$$\chi^2_{exp} = n\left[\sum_{i_1=1}^{K}\cdots\sum_{i_R=1}^{K}\frac{\left(\overline{p}_{i_1\ldots i_R} - \hat{p}_{i_1\ldots i_R}\right)^2}{\hat{p}_{i_1\ldots i_R}}\right] = n\left[\frac{1-\delta_{i_1\ldots i_R}}{1-\hat{\Delta}}\sum_{i_1=1}^{K}\cdots\sum_{i_R=1}^{K}\frac{\overline{p}^2_{i_1\ldots i_R}}{\prod_{r=1}^{r=R}\hat{\pi}_{i_r r}} - \overline{D}\right] \qquad \textbf{(12)}$$

*vs* $\chi^2_{\alpha,df}$, the (1−$\alpha$)-percentile of the chi-square distribution with $df=(K^R-1)-K-R(K-1)$ degrees of freedom. As shown in Appendix B, the second equality of expression (12) is attributed to the fact that the observed and expected proportions of agreements ($\overline{p}_i = \hat{p}_i$) and disagreements ($\overline{d}_{ir} = \hat{d}_{ir}$) are equal, which implies that the observed and expected marginal distributions are also equivalent ($\overline{t}_{ir} = \hat{t}_{ir}$). The value of $df$ is attributed to the fact that the multinomial distribution consists of ($K^R-1$) cells that are free to take values, from which we have to subtract $K$ parameters $\alpha_i$ estimated and $(K-1)R$ probabilities $\pi_{ir}$ estimated. The test will not be reliable due to the presence of many cells with very small expected proportions.

After obtaining the parameter estimates and measures of agreement of the *multi-rater delta* model ($\Delta$, $\alpha_i$, $\pi_{ir}$ and $\mathcal{S}_i$), their variances must be obtained to make inferences about the measures of agreement. Appendix C shows that

$$\hat{V}\left(\hat{\Delta}\right)=\frac{1-\hat{\Delta}}{n}\left\{\hat{\Delta}+\frac{\hat{X}}{\left(R-1\right)\hat{X}-1}\right\}, \tag{13}$$

$$\hat{V}\left(\hat{\alpha}_i\right)=\frac{1}{n}\left[\hat{\alpha}_i\left(1-\hat{\alpha}_i\right)+\left(1-\hat{\Delta}\right)\hat{X}_i\left\{\frac{\left(R-1\right)\hat{X}_i}{\left(R-1\right)\hat{X}-1}-1\right\}\right] \text{ and} \tag{14}$$

$$\hat{V}\left(\hat{\mathcal{S}}_i\right)=\frac{R^2}{n\bar{N}_i^2}\left[\begin{array}{l} n\hat{V}\left(\hat{\alpha}_i\right)-\hat{\alpha}_i\left(1-\hat{\alpha}_i\right)+\hat{\alpha}_i\left(1-\hat{\mathcal{S}}_i\right)\left\{1-\frac{R-1}{R}\hat{\mathcal{S}}_i\right\}+ \\ +\frac{\left(1-\hat{\Delta}\right)\hat{\mathcal{S}}_i^2}{R^2}\left\{\left(\sum_{r=1}^{R}\hat{\pi}_{ir}\right)^2-\left(\sum_{r=1}^{R}\hat{\pi}_{ir}^2\right)\right\}\end{array}\right] \tag{15}$$

where $\hat{X}_i=\left[\sum_{r=1}^{R}\hat{\pi}_{ir}^{-1}-\left(\prod_{r=1}^{R}\hat{\pi}_{ir}\right)^{-1}\right]^{-1}$ and $\hat{X}=\sum_{i=1}^{K}\hat{X}_i$. These estimated variances cannot be applied when any of the estimated parameters are at the boundary of the parametric space or are indeterminate. This condition occurs when $\bar{d}_{ir}$=0 or $B$=∞ because then there exists some $\hat{\pi}_{ir}$=0, as shown in *Supplementary Material*. In these cases, variances can be estimated if the calculations are carried out for the data increased by .5; thus, the new sample size is $n$+$K^R$/2, and the new proportions observed are $\bar{p}_{i_1i_2\ldots i_R}=\left(x_{i_1i_2\ldots i_R}+0.5\right)\big/\left(n+K^R/2\right)$.

As shown in *Supplementary Material*, all results obtained by the current *multi-rater delta* model are compatible with the results of the classic *delta* model, which is specifically defined for $R$=2 and $K$>2, with the exception of some very specific results in which errors were made by Martín Andrés & Femia Marzo (2004 and 2005). When one of these two raters is a gold-standard or when the marginal distribution is fixed beforehand, the classic *delta* model is preferred since it contemplates these two situations.

*4.2. Particular case with only two raters and only two categories (*$R$=$K$=2)

When only two raters and two categories exist, the problem with the *multi-rater delta* model is that there are more unknown parameters ($\alpha_1$, $\alpha_2$, $\pi_{11}$ and $\pi_{12}$) than free cells to take values (of which there are only three). In this case, the following solution by Martín Andrés & Femia Marzo (2004) can be adopted; it has been proved to provide coherent results (Martín

Andrés & Femia Marzo 2004, 2005 and 2008; Ato *et al.* 2011; Shankar and Bangdiwala 2014). The procedure is to create a third dummy category of observed frequencies $x_{i3}$=$x_{3j}$=0 ($\forall i, j$), increase all data in the new 3×3 table by .5, estimate the parameters as executed in the previous section, and redefine the measures of agreement without considering the third dummy category. In fact, Martín Andrés & Femia Marzo carried out their calculations for $x_{i3}$=$x_{3j}$=0 ($\forall i \neq j$) and $x_{33}$=1, but they themselves justified the fact that the value assigned to $x_{33}$ is irrelevant. Let $\bar{p}_i$ and $\bar{d}_{ir}$ be the new observed proportions and $\alpha_i$, $\pi_{ir}$ and $\Delta$ be the parameters of the *multi-rater delta* model; all parameters refer to the new 3×3 table. The measures of agreement for the original 2×2 table are defined as $\alpha_i^*=\alpha_i/(p_{1\bullet}+p_{2\bullet})$, $\Delta^* = \alpha_1^* + \alpha_2^*$ and $\mathcal{S}_i^*=2\alpha_i/(p_{1\bullet}+p_{2\bullet})$, for $i$=1 and 2; their estimates and estimated variances (see Appendix D) are

$$\hat{\alpha}_i^* = \frac{\hat{\alpha}_i}{1-\bar{p}_{3\bullet}},\ \hat{\mathrm{V}}\left(\hat{\alpha}_i^*\right) = \frac{1}{n\left(1-\bar{p}_{3\bullet}\right)^2}\left[\left(1-\hat{\Delta}\right)\hat{X}_i\left\{\frac{\hat{X}_i}{\hat{X}-1}-1\right\}+\left(1-\bar{p}_{3\bullet}\right)\hat{\alpha}_i^*\left(1-\hat{\alpha}_i^*\right)\right], \quad \textbf{(16)}$$

$$\hat{\Delta}^* = \hat{\alpha}_1^* + \hat{\alpha}_2^*,\ \hat{\mathrm{V}}\left(\hat{\Delta}^*\right) = \frac{1}{n\left(1-\bar{p}_{3\bullet}\right)^2}\left[\left(1-\hat{\Delta}\right)\left(1-\hat{X}_3\right)\frac{\hat{X}-\hat{X}_3}{\hat{X}-1}+\left(1-\bar{p}_{3\bullet}\right)\hat{\Delta}^*\left(1-\hat{\Delta}^*\right)\right], \quad \textbf{(17)}$$

$$\hat{\mathcal{S}}_i^* = \frac{2\hat{\alpha}_i}{\bar{N}_i},\ \hat{\mathrm{V}}\left(\hat{\mathcal{S}}_i^*\right) = \frac{4}{n\bar{N}_i^2}\left[\left(1-\hat{\Delta}\right)\hat{X}_i\left\{\frac{\hat{X}_i}{\hat{X}-1}-1\right\}+\hat{\alpha}_i\left\{1-\frac{3\hat{\alpha}_i}{\bar{N}_i}+\frac{2\hat{\alpha}_i^2}{\bar{N}_i^2}+\frac{2\hat{\pi}_{i1}\hat{\pi}_{i2}\left(1-\hat{\Delta}\right)\hat{\alpha}_i}{\bar{N}_i^2}\right\}\right], \quad \textbf{(18)}$$

where $\left(1-\bar{p}_{3\bullet}\right) = \bar{p}_1 + \bar{p}_2 + \bar{d}_{11} + \bar{d}_{21}$.

**5. Examples**

In this section, the data for the two examples described in section 1 will be analysed (as well as for two other examples which are a modification of the first two). As regards the data for the $R$=2 raters in Table 1(a), it has already been pointed out in section 2.1 that $\hat{\kappa}_C$=.6765. With only two raters, $\kappa_F$ is not usually calculated; its estimation $\hat{\kappa}_F$=.6753 is different from the value of Cohen' *kappa*. To apply the *multi-rater delta* model the first step is to construct Table 1(b), which is the appropriate table for this model; for example $\overline{p}_1 = \overline{p}_{11}$=.75, $\overline{d}_{11} = \overline{p}_{12} + \overline{p}_{13}$=.01+.04=.05, etc. Based on this table and the expression (11), the values $B$=.3125, $\lambda_1$=.05, $\lambda_2$=.00479 and $\lambda_3$=0 (which is directly derived from the fact that $\overline{d}_{31}$=0) are obtained. Finally, through expressions (10), the results of Table 1(c) are obtained. According to the *multi-rater delta* model, the degree of overall agreement $\hat{\Delta}$=.6875 (which is very close to $\hat{\kappa}_C$) is the sum of the proportion of agreements not due to chance obtained in each of the three classes ($\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3$=.5500+.0375+.1000), but the degree of agreement in each class is given by the consistencies $\hat{\mathcal{S}}_1 = .6875$, $\hat{\mathcal{S}}_2 = .5000$ and $\hat{\mathcal{S}}_3 = .8000$; hence class 3 is the one with the best degree of agreement between the two raters, while class 2 has the worst degree of agreement. These values are also quite close to those given by *kappa* when the data in Table 1(a) are collapsed in each of the three classes: $\hat{\kappa}_{C(1)}$=.6875, $\hat{\kappa}_{C(2)}$=.5000 y $\hat{\kappa}_{C(3)}$=.7727. According to Martín Andrés & Femia Marzo (2005) this similarity of results occurs when the two marginal distributions are homogenous (as in the example) and they are not excessively unbalanced. As regard the SE, they have all been obtained based on the data in Table 1(a) incremented by .5, because $\hat{\pi}_{31}$=0 (which is due to the fact that $\overline{d}_{31}$=0). In reality, the first step should be to verify the validity of the model; by applying the expression (12) to the data in Table 1(a), we obtain $\chi^2_{exp}$ = 0 ($df$=1) so that the *multi-rater delta* model is a perfect fit.

In the previous example the difference between $\hat{\kappa}_C$=.6765 and $\hat{\Delta}$=.6875 is small, but the situation changes when the marginals have even more imbalance. By modifying Table 1(a) as in Martín Andrés & Femia Marzo (2004, 2005), in which case the data by rows are 92/0/0 (row 1), 2/1/1 (row 2) and 2/1/1 (row 3), one obtains the values $\hat{\kappa}_C$=.479 and $\hat{\Delta}$=.920, which are now very different. The reason for this discrepancy is that the coefficient *kappa* does not take into account the information given by the marginal distributions, unlike coefficient *delta*. Note that in this example both raters are in agreement on classifying almost all the subjects into the psychotic type (92% and 96% respectively), a fact that *delta* takes into consideration but not *kappa*. In this example, once again $\chi^2_{exp} = 0$ (*df*=1).

As regards the data for the *R*=3 raters in Table 2(a), in order to apply the *multi-rater delta* model Table 2(b) must first be constructed; for example, $n\,\overline{p}_1$=56 and $n\,\overline{d}_{11}$= 1+0+5+3+0+0+0+1=10. Based on this table, and the expression (11), the values *B*=.450399427, $\lambda_1$=.010412048, $\lambda_2$=.077366093 and $\lambda_3$=.003015258 are obtained. Finally, the results in Table 2(c) are obtained from expressions (10). Note that the total agreement is 54.96% and the degrees of agreement in categories 1, 2 and 3 (consistencies) are 70.40%, 24.62%, and 63.06%, respectively. This finding indicates that category 2 has a very low consistency (24.62%), so the three raters should make efforts to homogenize their classification criteria especially in this category. These values are no longer so close to those given by Hubert's *kappa* when the data in Table 2(a) are collapsed in each of the three classes – $\hat{\kappa}_{HR(1)}$=63.62%, $\hat{\kappa}_{HR(2)}$=42.70% and $\hat{\kappa}_{HR(3)}$=60.81%–, especially in class 2; the reason is that now the marginal distributions of the three raters are not homogenous, due in particular to rater 2. The raw values of these four parameters, or values uncorrected for chance, are 100/164=60.98% for the total degree of agreement and 3×56/232=72.41%, 3×20/148=40.54%, and 3×24/112=64.29% for consistencies in categories 1, 2, and 3, respectively. The estimated total degree of agreement is .5496±.0462, which indicates that the real total degree of agreement is $\Delta \geq .5496-1.645\times.0462=.4736$ for 95% confidence. Finally, by applying the expression (12) to the data in Table 2(a), we obtain $\chi^2_{exp}$ = 155.41 (*df*=17); thus, the test is significant for a Type I error of $\alpha$=5%, and the *multi-rater delta* model does not fit the data. However, this result is not reliable because 7 (21) expected amounts less than one (less than or equal to five) exist, that is, 25.9% (77.8%); this is a normal occurrence when $R$>2. In addition, the raters can be evaluated to determine which of the three raters has the worst behaviour. Table 2(d) shows the degree of agreement parameters for all combinations of all raters save one. The total degree of agreement decreases when raters 1 or 2 are eliminated and increases when rater 3 is eliminated, indicating that rater 3 is the most divergent of the three raters. By eliminating raters 1 or 2, the consistency in category 2 becomes negative. The conclusion is that an effort should be made to homogenize the classification criterion, especially in the case of rater 3 and category 2.

Determining the different *kappa* measures of agreement requires summarizing the data of Table 2(a) in a different way. To calculate $\hat{\kappa}_{HR}$ one needs the data in Table 2(d), which are obtained from Table 2(b); based on this and the expression (5), $\hat{\kappa}_{HR}$=.5471. To calculate $\hat{\kappa}_F$ one needs the data in Table 2(f), which are obtained from Table 2(a); for example, the number 26 in Table 2(f), which is the number of subjects classified in category 1 by only one of the three raters, is obtained by summing all the frequencies $x_{i_1i_2i_3}$ of Table 2(a), in which on only one occasion $i_r$=1, that is, 26=3+1+2+1+14+1+2+2. Based on Table 2(f), $\sum_{s=1}^{164}\sum_{i=1}^{3} R_{si}^2$ = $72\times1^2+60\times2^2+100\times3^2$=1,212, $\sum_{i=1}^{3} R_i^2$ =$(232^2+148^2+112^2)/(164\times3)^2$=.364664 and, through the expression (6), $\hat{\kappa}_F$ =1−[164×9−1,212]/[164×3×2×(1−.364664)]=.5777. Finally, to calculate $\hat{\kappa}_{H2}$ the data in Tables 2(e) and (f) is needed; from the expression (5), $\hat{\kappa}_{H2}$=.5809. Table 2(g) summarizes the numerical results produced by applying the different measures of degree of agreement to the data in Table 2(a), where the raw degree of agreement is $\sum_{i=1}^{3} \overline{p}_i$ =(56+20+24)/164=.6098.

In the previous example, the differences between the various measures of the overall degree of agreement corrected for chance are small; however, the situation changes when the marginals are unbalanced. This is the case with the data in Table 3(a), which, as they are a modification of the data in Table 2(a), yield very unbalanced marginals. For example, the proportions of responses by rater 1 in categories 1, 2 and 3 are 115/164, 27/164, and 22/164, respectively. The results in Table 3(e) indicate that the *multi-rater delta* degree of agreement (70.8%) is slightly lower than the degree of raw agreement (74.4%), but significantly higher than the *kappa*-based agreement (57.4%, 55.5% y 55.4%), which is attributed to the fact that the *kappa* coefficients are influenced by the marginal distributions (because the *kappa* coefficients do not take into account the information provided by the marginal about the degree of agreement). Now $\chi^2_{exp}$ =19.83 ($df$=17) is not significant for $\alpha$=5%, even if 9 (24) expected amounts less than one (less than or equal to five) exist.

**6. Conclusions**

In this paper, we have looked at the evaluation of multi-rater agreement in the case of nominal categories. This issue is important in medicine, psychometry and whenever the intention has been to measure the degree of agreement among $R$ raters who classify $n$ subjects within $K$ categories.

When only two raters ($R$=2) exist, it is very common to use Cohen's *kappa* coefficient (1960) and, occasionally, the *delta* coefficient (Martín Andrés and Femia Marzo 2004). The *delta* coefficient has a three-fold advantage over the *kappa* coefficient. The first two advantages are that the *delta* agreement coefficient is not affected by imbalance in the marginal distributions of each rater and the *delta* model from which this coefficient proceeds has no difficulty in evaluating the degree of agreement in each category (Martín Andrés & Femia Marzo, 2005). The opposite occurs in the case of the *kappa* coefficient (Brennan & Prediger 1981; Agresti *et al.*, 1995; Guggenmoos-Holzmann & Vonk, 1998). The third advantage is that the *delta* model consists of specific parameters for evaluating the agreement in each category when one of the raters is a gold-standard.

When many raters ($R$≥2) exist, several *kappa*-based coefficients can be employed, such as Fleiss' *kappa* (1971), Hubert's *kappa* (1977), and Conger's *kappa* (1980). However, Fleiss' *kappa* does not coincide with Cohen's *kappa* when $R$=2. The defects of all multi-rater *kappa* coefficients are identical to the defects of the Cohen's *kappa* coefficient (Marasini *et al.*, 2016; Conger, 2017). In this article, the *delta* model has been modified to ensure that it is valid in the multi-rater case; thus, the *multi-rater delta* model is obtained. This model assumes that none of the raters is a gold-standard and provides results that are compatible with the results of the classic *delta* model. In addition, the parameters provided by the *multi-rater delta* model have two advantages over the multi-rater *kappa* coefficients. In the first place, the degree of total agreement of the *multi-rater delta* model (parameter $\Delta$) is not affected by the marginal distributions of the raters, unlike the multi-rater *kappa* coefficients. Second, the *multi-rater delta* model allows the degree of agreement in each class to be measured through the concept of consistency (coefficients $\mathcal{S}_i$). However, multi-rater *kappa* coefficients usually measure the degree of agreement in each class by collapsing the data in this class, this means that the collapsed *kappa* coefficients attempt to measure simultaneously the degree of agreement in this class, the degree of agreement in all the remaining classes as a single class and the whole degree of agreement in the collapsed table. It is not possible to reconcile these three objectives simultaneously. A program (Multi-Rater Delta) to run the *multi-rater delta* model can be downloaded free of charge at http://www.ugr.es/local/bioest/software (Section "Agreement Among Raters").

**7. Acknowledgements**

This research was supported by the Spanish Ministry of the Economy, Industry and Competitiveness under grant number MTM2016-76938-P (co-financed by funding from FEDER). The authors declare that there is no conflict of interest. The authors wish to thank Dr. Pedro Femia Marzo for his invaluable assistance in various aspects of this article.

# APPENDICES

## A. Relationships between the parameters of two (classic and new) *delta* models for *R*=2

Under the *delta* model given by expression (1), Martín Andrés & Femia Marzo (2004, 2005) defined the following three parameters for the case of any two raters

$$\mathcal{A}_i = p_{i\bullet}\Delta_i,\ \Delta=\Sigma\,\mathcal{A}_i \text{ and } \mathcal{S}_i=2p_{i\bullet}\Delta_i/(p_{i\bullet}+p_{\bullet i}) \qquad \textbf{(A1)}$$

where $\mathcal{A}_i$ is the proportion of the agreements in category *i* that do not occur by chance; $\Delta$ is the overall proportion of agreements that are not due to chance (that is, the overall degree of agreement); and $\mathcal{S}_i$ is the *consistency* in category *i*, which measures the degree of agreement in category *i*. Due to expressions (1) and (A1),

$$p_{ij}=\delta_{ij}\alpha_i+(p_{i\bullet}-\alpha_i)\pi_j \text{ with } \alpha_i=p_{i\bullet}\Delta_i=\mathcal{A}_i \qquad \textbf{(A2)}$$

If $\Delta=\sum_{i=1}^{K}\alpha_i=\sum_{i=1}^{K}p_{i\bullet}\Delta_i$, from expression (A2), $p_{\bullet j}=\sum_{i=1}^{K}p_{ij}=p_{jj}+\sum_{i=1,i\neq j}^{K}p_{ij}=\alpha_j+(p_{j\bullet}-\alpha_j)\pi_j+\pi_j\sum_{i=1,i\neq j}^{K}\left(p_{i\bullet}-\alpha_i\right)=\alpha_j+\pi_j\sum_{i=1}^{K}\left(p_{i\bullet}-\alpha_i\right)=\alpha_j+(1-\sum_{i=1}^{K}\alpha_i)\pi_j=\alpha_j+(1-\Delta)\pi_j$.

The *delta* model of expression (1) is expressed "in rows", that is, taking rater 1 as the reference. If the model is "in columns", that is, taking rater 2 as the reference, it will depend on new parameters $\tilde{\Delta}_j$ and $\tilde{\pi}_j$, similar to $\Delta_i$ and $\pi_i$ as previously defined. Following the reasoning in the previous paragraph, $p_{ij}=\delta_{ij}\tilde{\alpha}_j+\left(p_{\bullet j}-\tilde{\alpha}_j\right)\tilde{\pi}_i$, with $\tilde{\alpha}_j=p_{\bullet j}\tilde{\Delta}_j$, $p_{i\bullet}=\tilde{\alpha}_i+\left(1-\tilde{\Delta}\right)\tilde{\pi}_i$ and $\tilde{\Delta}=\sum_{j=1}^{K}\tilde{\alpha}_j$. Since $p_{ij}$ must take the same value in both models, $p_{ij}=(p_{i\bullet}-\alpha_i)\pi_j=\left(p_{\bullet j}-\tilde{\alpha}_j\right)\tilde{\pi}_i$ for $i\neq j$, that is, $(p_{i\bullet}-\alpha_i)/\tilde{\pi}_i=\left(p_{\bullet j}-\tilde{\alpha}_j\right)/\pi_j$ ($\forall i\neq j$). Assuming that $K>2$, since the case of $K=2$ requires special treatment, the previous equality takes the constant value $\gamma$, and so

$$p_{i\bullet}-\alpha_i=\gamma\tilde{\pi}_i \text{ and } p_{\bullet j}-\tilde{\alpha}_j=\gamma\pi_j \qquad \textbf{(A3)}$$

By adding the values in *i* or *j*, we obtain $1-\sum_{i=1}^{K}\alpha_i=1-\sum_{i=1}^{K}\tilde{\alpha}_i=\gamma$, and $\Delta=\tilde{\Delta}$. In the case of

$i$=$j$, $p_{ii}$=$\alpha_i$+($p_{i\bullet}$−$\alpha_i$)$\pi_i$=$\tilde{\alpha}_i + \left(p_{\bullet i} - \tilde{\alpha}_i\right)\tilde{\pi}_i$, so $\alpha_i$+$\gamma\pi_i\tilde{\pi}_i = \tilde{\alpha}_i + \gamma\tilde{\pi}_i\pi_i$ and $\alpha_i$=$\tilde{\alpha}_i$. Substituting expression (A3) into (A2) and considering that $\gamma$=1−$\Delta$, we obtain $p_{ij}$=$\delta_{ij}\alpha_i$+(1−$\Delta$)$\tilde{\pi}_i\pi_j$, which is the *multi-rater delta* model of expression (2). The result is that it is irrelevant whether the *delta* model is expressed "in rows" or "in columns" because the parameters of expression (A1) are equivalent in both cases.

The values obtained in the previous paragraphs yield the parameters of the *delta* model defined "in columns" according to the parameters of the *delta* model defined "in rows":

$$p_{\bullet i} = p_{i\bullet}\Delta_i + (1-\Delta)\pi_i,\ \tilde{\Delta}_i = \frac{p_{i\bullet}\Delta_i}{p_{i\bullet}\Delta_i + (1-\Delta)\pi_i} \text{ and } \tilde{\pi}_i = \frac{p_{i\bullet}(1-\Delta_i)}{1-\Delta} \quad \textbf{(A4)}$$

The first equality stems from the final statement in the first paragraph; the second equality from the fact that $\alpha_i$=$p_{i\bullet}\Delta_i$=$\tilde{\alpha}_i = p_{\bullet i}\tilde{\Delta}_i$; and the third equality from the fact that $\tilde{\pi}_i$=($p_{i\bullet}$−$\alpha_i$)/$\gamma$= ($p_{i\bullet}$−$\tilde{\alpha}_i$)/$\gamma$=$p_{i\bullet}$(1−$\Delta_i$)/(1−$\Delta$). Based on all the foregoing, the following relationships between the parameters of the new *delta* model (section 2.3) and of the classic *delta* model (section 2.2), as defined “in rows” can also be deduced.

$$\alpha_i = p_{i\bullet}\Delta_i,\quad \Delta=\sum\nolimits_{i=1}^{K} p_{i\bullet}\Delta_i,\quad \pi_{i2}=\pi_i,\ \pi_{i1}=p_{i\bullet}(1-\Delta_i)/(1-\Delta),\quad p_{i\bullet}=\alpha_i+(1-\Delta)\pi_{i1} \quad \textbf{(A5)}$$

**B. Maximum likelihood estimators of the *multi-rater delta* model when *R*>2 or *K*>2**

To simplify this explanation, in the following proofs will be provided when $R$=3; these can be extended directly to any other case, except in the case of certain aspects that will be specified as they arise. In addition, now the *multi-rater delta* model of expression (7) is simplified to $p_{ijh}$=$\delta_{ijh}\alpha_i$+(1−$\Delta$)$\pi_{i1}\pi_{j2}\pi_{h3}$ with $i$, $j$, $h$=1, 2, …,$K$ and $\Delta$=$\sum\nolimits_{i=1}^{K}\alpha_i$. If $\overline{p}_{ijh}$=$x_{ijh}/n$ are the observed proportions, then, except for one constant the logarithm of the likelihood is $L$= $\sum\nolimits_{i=1}^{K}\sum\nolimits_{j=1}^{K}\sum\nolimits_{h=1}^{K}\overline{p}_{ijh}\log p_{ijh} = \sum\nolimits_{i=1}^{K}\overline{p}_{iii}\log p_{iii} + \sum\nolimits_{i=1}^{K}\sum\nolimits_{j=1}^{K}\sum\nolimits_{h=1}^{K}\overline{p}_{ijh}\,(1-\delta_{ijh})\log p_{ijh} = \sum\nolimits_{i=1}^{K}\overline{p}_{iii}\log p_{iii} +$ $\sum\nolimits_{i=1}^{K}\sum\nolimits_{j=1}^{K}\sum\nolimits_{h=1}^{K}\overline{p}_{ijh}\left(1-\delta_{ijh}\right)\left\{\log(1-\Delta)+\log\pi_{i1}+\log\pi_{j2}+\log\pi_{h3}\right\}$. By working it out we obtain $L=\sum_{i=1}^{K}\bar{p}_{iii}\log p_{iii}+(1-\bar{p})\log(1-\Delta)+\sum_{i=1}^{K}\sum_{r=1}^{R}(\log\pi_{ir})\bar{d}_{ir}$, where $\bar{p}=\sum_{i=1}^{K}\bar{p}_{iii}$ is the total proportion of agreements, $r$=1, 2, 3 and $\bar{d}_{ir}$ is the proportion of disagreements in category $i$ of rater $r$, that is, $\bar{d}_{i1}=\bar{p}_{i\bullet\bullet}-\bar{p}_{iii}$, $\bar{d}_{i2}=\bar{p}_{\bullet i\bullet}-\bar{p}_{iii}$ and $\bar{d}_{i3}=\bar{p}_{\bullet\bullet i}-\bar{p}_{iii}$.

Bearing in mind that $p_{iii}=\alpha_i+(1-\Delta)\pi_{i1}\pi_{i2}\pi_{i3}$, the following derivatives are obtained

$$\frac{\partial p_{iii}}{\partial\alpha_t}=\delta_{it}-\frac{p_{iii}-\alpha_i}{1-\Delta},\ \frac{\partial p_{iii}}{\partial\pi_{tr}}=\delta_{it}\frac{p_{iii}-\alpha_i}{\pi_{ir}},$$

and the derivatives of $L$ are

$$\frac{dL}{d\alpha_i}=\frac{\bar{p}_{iii}}{p_{iii}}-\frac{A}{1-\Delta}\text{ with }A=1-\sum_{i=1}^{K}\frac{\bar{p}_{iii}}{p_{iii}}\alpha_i,\ \frac{\partial L}{\partial\pi_{ir}}=\frac{1}{\pi_{ir}}\left\{\bar{d}_{ir}+\bar{p}_{iii}-\frac{\bar{p}_{iii}}{p_{iii}}\alpha_i\right\}=A_{ir} \qquad \textbf{(B1)}$$

Since the maximum likelihood estimates of $\alpha_i$ must verify that $dL/d\alpha_i$=0, then $\bar{p}_{iii}/p_{iii}=A/(1-\Delta)$ through the first expression of (B1). By substituting in the definition of $A$, $A=1-\Delta$, $\bar{p}_{iii}/p_{iii}=1$ and $\bar{p}_{iii}=p_{iii}$. The last equality indicates that if the maximum likelihood estimates $\hat{\alpha}_i$ and $\hat{\pi}_{ir}$ of $\alpha_i$ and $\pi_{ir}$, respectively, were known, the maximum likelihood estimator for $p_{iii}$ would be $\hat{p}_{iii}=\hat{\alpha}_i+(1-\hat{\Delta})\hat{\pi}_{i1}\hat{\pi}_{i2}\hat{\pi}_{i3}=\bar{p}_{iii}$. Thus

$$\hat{\alpha}_i=\bar{p}_{iii}-(1-\hat{\Delta})\hat{\pi}_{i1}\hat{\pi}_{i2}\hat{\pi}_{i3}\text{ and }\hat{p}_{iii}=\bar{p}_{iii}, \qquad \textbf{(B2)}$$

where $\hat{\Delta}=\sum_{i=1}^{K}\hat{\alpha}_i$ is the maximum likelihood estimator of $\Delta$.

Substituting $\bar{p}_{iii}=p_{iii}$ in the second expression of (B1), $\partial L/\partial\pi_{ir}=(\bar{d}_{ir}+\bar{p}_{iii}-\alpha_i)/\pi_{ir}=A_{ir}$. If the maximum likelihood estimators of $\pi_{ir}$ have to verify that $dL/d\pi_{ir}$=0 and $\sum_{i=1}^{K}\pi_{ir}=1$, then $\partial L/\partial\pi_{ir}=\partial L/\partial\pi_{jr}$, $A_{ir}=A_{jr}$ ($\forall i, j$), $A_{ir}=A_r$ and $\sum_{i=1}^{K}A_r\pi_{ir}=A_r=1-\Delta$. Therefore, $A_r=A=1-\Delta$. Given that $\bar{p}_{iii}-\alpha_i=p_{iii}-\alpha_i=(1-\Delta)\pi_{i2}\pi\pi_{i3}$, from the last expression (B1), we obtain

$$\frac{\partial L}{\partial \pi_{ir}} = \frac{1}{\pi_{ir}}\left\{\bar{d}_{ir} + (1-\alpha)\pi_{i1}\pi_{i2}\pi_{i3}\right\} = 1-\mathit{\Delta} \quad \text{or} \quad (1-\mathit{\Delta})\pi_{ir} = \bar{d}_{ir} + (1-\mathit{\Delta})\pi_{i1}\pi_{i2}\pi_{i3}. \qquad \textbf{(B3)}$$

These expressions yield the maximum likelihood estimators $\hat{\mathit{\Delta}}$ and $\hat{\pi}_{ir}$. Once the estimators are obtained, $\hat{d}_{i1} = \hat{p}_{i\bullet\bullet} - \hat{p}_{iii} = (1-\hat{\mathit{\Delta}})\hat{\pi}_{i1} - (1-\hat{\mathit{\Delta}})\hat{\pi}_{i1}\hat{\pi}_{i2}\hat{\pi}_{i3}$; thus, $\hat{d}_{i1} = \bar{d}_{i1}$ from the last equality of (B3) and $\hat{p}_{i\bullet\bullet} = \bar{p}_{i\bullet\bullet}$. In general, $\hat{d}_{ir} = \bar{d}_{ir}$, $\hat{p}_{i\bullet\bullet} = \bar{p}_{i\bullet\bullet}$, $\hat{p}_{\bullet i\bullet} = \bar{p}_{\bullet i\bullet}$ and $\hat{p}_{\bullet\bullet i} = \bar{p}_{\bullet\bullet i}$; and as it was previously indicated that $\bar{p}_{iii} = \hat{p}_{iii}$, then $\bar{t}_{ir} = \hat{t}_{ir}$.

The estimators of $\mathit{\Delta}$ and $\pi_{ir}$ can be obtained based on the expressions (B3). If $\bar{d}_{i3}$=0, then $\partial L/\partial\pi_{i3}$=$(1-\mathit{\Delta})\pi_{i1}\pi_{i2} \leq 1-\mathit{\Delta}$, $\mathrm{d}L/\mathrm{d}\pi_{i3} \leq 0$, and $\hat{\pi}_{i3}$=0; generally if $\bar{d}_{ir}$=0, then $\hat{\pi}_{ir}$=0. Let $\lambda_i$=$(1-\mathit{\Delta})\pi_{i1}\pi_{i2}\pi_{i3}$; the last equality in expression (B3) indicates that $\pi_{ir}$=$(\bar{d}_{ir}+\lambda_i)/(1-\mathit{\Delta})$, and therefore, $\lambda_i$=$(\bar{d}_{i1}+\lambda_i)$ $(\bar{d}_{i2}+\lambda_i)$ $(\bar{d}_{i3}+\lambda_i)/(1-\mathit{\Delta})^2$ ($\forall i$). If $B$=$1-\mathit{\Delta}$, the following expressions must be solved for $\lambda_i$:

$$B^2 = \frac{\prod_{r=1}^{3}\left(\lambda_i + \bar{d}_{ir}\right)}{\lambda_i} \ (\forall i \mid \bar{d}_{ir} \neq 0, \forall r) \text{ under the condition } g(B) = \sum_{i=1}^{K}\lambda_i - B + \bar{D} = 0 \ \textbf{(B4)}$$

The condition $g(B)$=0 is due to the fact that when $\pi_{ir}$=$(\bar{d}_{ir}+\lambda_i)/B$ and $\sum_{i=1}^{K}\pi_{ir}$=1, then $B\sum_{i=1}^{K}\pi_{ir} = \bar{D} + \sum_{i=1}^{K}\lambda_i$=$B$. Expression (B4) is generalized in expression (11); in *Supplementary Material* it is shown how to obtain the solutions to this expression. Once the values of $B$ and $\lambda_i$ are determined, taking into account that $\pi_{ir}$=$(\bar{d}_{ir}+\lambda_i)/B$, the first equality of (B2), the definition of (8), and the fact that the observed and expected marginal distributions are equal, the following estimators are generalized in expression (10):

$$\hat{\pi}_{ir} = \frac{\lambda_i + \bar{d}_{ir}}{B},\ \hat{\alpha}_i = \bar{p}_{iii} - \lambda_i,\ \hat{\mathit{\Delta}} = 1-B \text{ and } \hat{\mathcal{S}}_i = \frac{3\hat{\alpha}_i}{\bar{p}_{i\bullet\bullet} + \bar{p}_{\bullet i\bullet} + \bar{p}_{\bullet\bullet i}} = \frac{3\hat{\alpha}_i}{3\bar{p}_i + \bar{D}_i}, \qquad \textbf{(B5)}$$

Note that by substituting $\lambda_i + \bar{d}_{ir}$=$B\hat{\pi}_{ir}$ in the first expression (11) we obtain

$$B = \frac{\bar{d}_{ir}}{\hat{\pi}_{ir} - \prod_{r'=1}^{R} \hat{\pi}_{ir'}} \quad (\forall i, r), \tag{B6}$$

so that the values of $\hat{\pi}_{ir}$ depend solely on the values of $\bar{d}_{ir}$ and not on the values of $\bar{t}_{ir}$.

By reason of expression (9)

$$\hat{\Delta} = \frac{\bar{I}_o - \hat{I}_\pi}{1 - \hat{I}_\pi}, \text{ where } \hat{I}_\pi = \sum_{i=1}^{K} \Pi_{r=1}^{R} \hat{\pi}_{ir},$$

and through the expression (5),

$$\hat{\Delta} - \hat{\kappa}_{HR} = \frac{\left(1 - \hat{\Delta}\right)\left(1 - \hat{\kappa}_{HR}\right)}{\left(1 - \bar{I}_e\right)} \left(\bar{I}_e - \hat{I}_\pi\right), \tag{B7}$$

so that $\hat{\Delta} - \hat{\kappa}_{HR}$ is proportional to $\bar{I}_e - \hat{I}_\pi$. The maximum (1) and minimum (0) values of $\bar{I}_e$ are reached respectively when $\bar{t}_{i1} = \bar{t}_{i2} = ... = \bar{t}_{iR} = 1$ in any $i$ (in which case all the remaining $\bar{t}_{jr}$=0) or when $\{\bar{t}_{i1}, \bar{t}_{i2}, ..., \bar{t}_{iR}\}$ in all the sets, one of the terms has a value of 1 and the rest have a value of 0. In the same way for $\hat{I}_\pi$, but now based on the probabilities $\hat{\pi}_{ir}$. In most of the practical situations the values of $\bar{t}_{ir}$ and $\hat{\pi}_{ir}$ are not close to these extreme cases, so that the value of $\bar{I}_e - \hat{I}_\pi$ is not excessive and the value of $\hat{\Delta} - \hat{\kappa}_{HR}$ will not be either. Usually, the marginal distributions $\bar{t}_{ir}$ are homogenous; if they were totally homogenous then $\bar{t}_{ir} = \bar{t}_i$ ($\forall r$), that is $\bar{d}_{ir} = \bar{d}_i$ ($\forall r$) and, through expression (B6), $\hat{\pi}_{ir} = \hat{\pi}_i$ ($\forall r$), where

$$B = \frac{\bar{d}_i}{\hat{\pi}_i - \hat{\pi}_i^R} \quad (\forall i) \text{ and } \sum_{i=1}^{K} \bar{t}_i = \sum_{i=1}^{K} \hat{\pi}_i = 1. \tag{B8}$$

The result of this is that $\bar{I}_e = \sum_{i=1}^{K} \bar{t}_i^R$ and $\hat{I}_\pi = \sum_{i=1}^{K} \hat{\pi}_i^R$. In these cases, the minimum (or maximum) value of $\bar{I}_e$ is $1/K^{R-1}$ (or 1), which is reached when $\bar{t}_i$=1/$K$ (or when all the $\bar{t}_i$ have a value of 0, except one which has a value of 1); it is similar with $\hat{I}_\pi$, but now based on the probabilities $\hat{\pi}_i$. When the marginal distributions of the proportions of responses are totally balanced, $\overline{t}_i$=1/$K$ (∀$i$), $\overline{I}_e$ takes the minimum value and $\hat{\Delta} \leq \hat{\kappa}_{HR}$ through expression (B7). When the marginal distributions of the proportions of disagreements are totally balanced, $\overline{d}_i = \overline{D}/K$ (∀$i$), $B = \overline{D}/K\left\{\hat{\pi}_i - \hat{\pi}_i^R\right\}$ (∀$i$), $\hat{\pi}_i$=1/$K$ (∀$i$), $\hat{I}_\pi$ takes the minimum value and $\hat{\Delta} \geq \hat{\kappa}_{HR}$ through expression (B7). However, as was previously pointed out, the important differences between $\hat{\Delta}$ and $\hat{\kappa}_{HR}$ occur when $\overline{I}_e$ approaches its maximum value. If $\overline{t}_1$ is very large, the rest of the values of $\overline{t}_i$ will be small and $\overline{I}_e$ will be close to 1; but that large marginal unbalance does not imply that the values $\overline{d}_i$ are also very unbalanced, so that, through expression (B8), the values of $\hat{\pi}_i$ will not be very unbalanced, $\hat{I}_\pi$ will not be close to its maximum value (which is 1), $\hat{I}_\pi$ will be much smaller that $\overline{I}_e$ and, finally, $\hat{\Delta}$ will be appreciably larger than $\hat{\kappa}_{HR}$.

**C. Variances of estimates when *R*>2 or *K*>2**

Agresti (2013) specifies the multivariate delta method in the case of a multinomial. If a function $f$=$f$($\boldsymbol{p_{ijh}}$) is estimated by $\hat{f} = f\left(\hat{\boldsymbol{p}}_{\boldsymbol{ijh}}\right)$, with $\boldsymbol{p_{ijh}}$={$p_{ijh}$}, $\hat{\boldsymbol{p}}_{\boldsymbol{ijh}} = \left\{\hat{p}_{ijh}\right\}$ and $\hat{p}_{ijh}$ the maximum likelihood estimates of $p_{ijh}$, then if $f_{ijh} = \partial f/\partial p_{ijh}$, the variance of $\hat{f}$ is

$$V\left(\hat{f}\right) = \left[\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K} f_{ijh}^2 p_{ijh} - \left(\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K} f_{ijh}\right)^2\right] \Big/ n. \quad \textbf{(C1)}$$

A similar notation to the one used for the observed proportions ($\overline{p}_{ijh}, \overline{d}_{ir}, \overline{p}_i, \overline{t}_{ir},$) will be applied for the different probabilities ($p_{ijh}$, $d_{ir}$, $p_i$, and $t_{ir}$). Any condition that an estimator must verify with respect to the observed proportions should also occur with respect to the probabilities. Since $\lambda_i/(\lambda_i+d_{ir})=B\pi_{i1}\pi_{i2}\pi_{i3}/B\pi_{ir}$, then

$$\frac{\lambda_i}{\lambda_i + d_{ir}} = \frac{P_i}{\pi_{ir}} \text{ and } \sum_{r=1}^{R}\frac{\lambda_i}{\lambda_i + d_{ir}} = P_i I_i \text{ with } P_i = \prod_{r=1}^{R}\pi_{ir} \text{ and } I_i = \sum_{r=1}^{R}\pi_{ir}^{-1}. \quad \textbf{(C2)}$$

Let $X_i$=$P_i$/($P_iI_i$–1)=1/($I_i$– $P_i^{-1}$) and $X$=$\sum_{i=1}^{K} X_i$.

When $\hat{f} = 1 - \hat{\Delta}$, we have $f$=$B$ and $V(B) = V(\hat{\Delta})$, with $B$=1−$\Delta$ given by expression (11), which is expressed in terms of probabilities. Because $g(B)$=0 and $\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K} p_{ijh} = 1$, 0=d$g$/d$p_{ijh}$=$\partial g/\partial p_{ijh} - \partial g/\partial p_{KKK} = \partial g/\partial p_{ijh}$ since $p_{KKK}$ does not intervene in $g$. Therefore, $0 = \partial g/\partial p_{ijh} = (\partial g/\partial B)(\partial B/\partial p_{ijh}) + (\partial g/\partial p_{ijh}) = -\Delta'_{ijh}[\Sigma_t \lambda'_{t(B)} - 1] + [\lambda'_{t(ijh)} + D'_{ijh}]$, with $\Delta'_{ijh} = \partial\Delta/\partial p_{ijh}$, $\lambda'_{t(B)} = \partial\lambda_t/\partial B$, $\lambda'_{t(ijh)} = \partial\lambda_t/\partial p_{ijh}$ and $D'_{ijh} = \partial D/\partial p_{ijh}$. Therefore,

$$\Delta'_{ijh} = \frac{\sum_{t=1}^{K} \lambda'_{t(ijh)} + D'_{ijh}}{\sum_{t=1}^{K} \lambda'_{t(B)} - 1}. \qquad \textbf{(C3)}$$

For the first equality of (11), 2log $B$=$\sum_{r=1}^{R} \log(\lambda_t + d_{tr})$−log $\lambda_t$; deriving this expression, we obtain $\lambda'_{t(B)}$=2$X_t$, and the denominator of expression (C3) will be 2$X$−1. To calculate the numerator, we need $D'_{ijh}$=1−$\delta_{ijh}$ and $\lambda'_{t(ijh)}$. If $i$=$j$=$h$, then $\lambda'_{t(ijh)}$=0 since the first equality of (11) does not depend on $p_{iii}$. Deriving this equality with respect to $p_{ijh}$, we obtain $\lambda'_{t(ijh)}[1 - P_t I_t] = \sum_{r=1}^{R} (P_t/\pi_{tr})(\partial d_{tr}/\partial p_{ijh})$ and $\lambda'_{t(ijh)} = -(1-\delta_{ijh}) X_t\{(\delta_{ti}/\pi_{t1}) + (\delta_{tj}/\pi_{t2}) + (\delta_{th}/\pi_{t3})\}$; therefore, $\sum_{t=1}^{K} \lambda'_{t(ijh)} = \lambda'_{i(ijh)} + \lambda'_{j(ijh)} + \lambda'_{h(ijh)} = -(1-\delta_{ijh})[1 - \{(X_i/\pi_{i1}) + (X_j/\pi_{j2}) + (X_h/\pi_{h3})\}]$. Replacing everything in expression (C3),

$$\Delta'_{ijh} = f_{ijh} = \frac{1-\delta_{ijh}}{2X-1}\left[1 - \left\{\frac{X_i}{\pi_{i1}} + \frac{X_j}{\pi_{j2}} + \frac{X_h}{\pi_{h3}}\right\}\right], \qquad \textbf{(C4)}$$

where the number 2 will be ($R$−1). To apply expression (C1), we must consider that $(1-\delta_{ijh})p_{ijh} = (1-\delta_{ijh})(\delta_{ijh}\alpha_i + B\pi_{i1}\pi_{j2}\pi_{h3}) = (1-\delta_{ijh})B\pi_{i1}\pi_{j2}\pi_{h3}$, $\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K} f_{ijh} p_{ijh} = -B$, $\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K} f_{ijh}^2 p_{ijh} = B(3X-1)/(2X-1)$ -which in general will be $B(RX-1)/[(R-1)X-1]$- and

$$V(\hat{\Delta}) = \frac{1-\Delta}{n}\left\{\Delta + \frac{X}{(R-1)X-1}\right\}. \qquad \textbf{(C5)}$$

If $\hat{f}=\hat{\alpha}_t$ then $f$=$p_{ttt}$−$\lambda_t$ and $\Delta'_{t(ijh)}$=$f_{ijh}$=$\partial\Delta/\partial p_{ijh}$=$\delta_{ijh}$−$\lambda'_{t(ijh)}+\lambda'_{t(B)}\Delta'_{ijh}$. Substituting the derivatives, which are known from the previous paragraph, for $R$=3,

$$\Delta'_{t(ijh)}=\delta_{ijht}+\left(1-\delta_{ijh}\right)X_t\left[\left\{\frac{\delta_{it}}{\pi_{t1}}+\frac{\delta_{jt}}{\pi_{t2}}+\frac{\delta_{ht}}{\pi_{t3}}\right\}+\frac{(R-1)}{(R-1)X-1}\left\{1-\left(\frac{X_i}{\pi_{i1}}+\frac{X_j}{\pi_{j2}}+\frac{X_h}{\pi_{h3}}\right)\right\}\right].\ \textbf{(C6)}$$

By performing the necessary operations and generalizing this to any value of $R$, we obtain $\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K}f_{ijh}p_{ijh}=\alpha_t$, $\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K}f_{ijh}^2p_{ijh}=\alpha_t+BX_t[(R-1)X_t/\{(R-1)X-1\}-1]$. Through expression (C1), we obtain

$$V\left(\hat{\alpha}_t\right)=\frac{1}{n}\left[\alpha_t\left(1-\alpha_t\right)+\left(1-\Delta\right)X_t\left\{\frac{(R-1)X_t}{(R-1)X-1}-1\right\}\right].\qquad\textbf{(C7)}$$

Finally, if $\hat{f}=\hat{\mathcal{S}}_t$ then, by expression (B5), $f$=$\mathcal{S}_t$=3$\alpha_t/N_t$ and $f_{ijh}$=$\partial\mathcal{S}_t/\partial p_{ijh}$={3/$N_t$}× [$\Delta'_{t(ijh)}$−$\alpha_t$($\delta_{ti}$+$\delta_{tj}$+$\delta_{th}$)/$N_t$] with $N_t$=$p_{t\bullet\bullet}$+$p_{\bullet t\bullet}$+$p_{\bullet\bullet t}$=3$p_t$+$D_t$. We obtain $\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K}f_{ijh}p_{ijh}=0$ and $\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{h=1}^{K}f_{ijh}^2p_{ijh}=n\mathrm{V}(\hat{\mathcal{S}}_t)$ by expression (C1), with

$$V\left(\hat{\mathcal{S}}_t\right)=\frac{1}{n}\left(\frac{3}{N_t}\right)^2\left[\left\{\frac{2X_t}{2X-1}-1\right\}BX_t+\alpha_t-\frac{(2\times3-1)\alpha_t^2}{N_t}+2\left(\frac{\alpha_t}{N_t}\right)^2\left(p_{tt\bullet}+p_{t\bullet t}+p_{\bullet tt}\right)\right].\ \textbf{(C8)}$$

Generalizing to any value of $R$, $N_t$=$Rp_t$+$D_t$=$R\alpha_i$+(1−$\Delta$)$\sum_{r=1}^{R}\pi_{ir}$, and

$$V\left(\hat{\mathcal{S}}_t\right)=\frac{R^2}{nN_t^2}\left[\begin{array}{l}BX_t\left\{\frac{(R-1)X_t}{(R-1)X-1}-1\right\}+\alpha_t\left(1-\mathcal{S}_t\right)\left\{1-\frac{R-1}{R}\mathcal{S}_t\right\}+\\+B\left(\frac{\mathcal{S}_t}{R}\right)^2\left\{\left(\sum_{r=1}^{R}\pi_{tr}\right)^2-\left(\sum_{r=1}^{R}\pi_{tr}^2\right)\right\}\end{array}\right].\qquad\textbf{(C9)}$$

The variances obtained in expressions (C5), (C7) and (C9) must be estimated based on the data of the problem; hence the expressions (13), (14) and (15), respectively.

**D. Variances in the particular case of *R*=*K*=2**

As indicated in section 4.2, the observed proportions verify the equalities

$\overline{p}_{i3} = \overline{p}_{3j} = \overline{p}_{33}$; thus, $\overline{d}_{11} = \overline{d}_{22}$, $\overline{d}_{12} = \overline{d}_{21}$ and $\overline{d}_{31} = \overline{d}_{32} = 2\overline{p}_{33}$. If we take the first expression (10), we obtain $\hat{\pi}_{11} = \hat{\pi}_{22}$, $\hat{\pi}_{12} = \hat{\pi}_{21}$ and $\hat{\pi}_{31} = \hat{\pi}_{32}$, so $\hat{X}_1 = \hat{X}_2$ and $\hat{X}_3 = \hat{\pi}_{31}^2 / \left(2\hat{\pi}_{31} - 1\right)$. Determining $\hat{V}\left(\hat{\mathcal{S}}_i\right)$ does not pose any problem, as the consistency $\hat{\mathcal{S}}_i$ in this case is defined in the same manner as it is defined in section 4.1. Therefore, its variance is similar to the variance of expression (15) for $R$=2, that is, the variance of expression (18).

If $\hat{f} = \hat{\alpha}_t^*$, then $f = \alpha_t^* = \alpha_t/(1-p_{3\bullet})$ and $f_{ij} = (\partial f/\partial \alpha_t)\times(\partial \alpha_t/\partial p_{ij}) + (\partial f/\partial p_{3\bullet})(\partial p_{3\bullet}/\partial p_{ij}) = [\partial \alpha_t/\partial p_{ij} + \alpha_t^* \delta_{i3}]/(1-p_{3\bullet})$, with $\partial \alpha_t/\partial p_{ij} = \Delta'_{t(ij)}$ given by expression (C6) for $R$=2 and omitting the subscript and terms in $h$. By performing the operations $\sum_{i=1}^{K}\sum_{j=1}^{K} f_{ij} p_{ij} = \alpha_t^*/(1-p_{3\bullet})$, $\sum_{i=1}^{K}\sum_{j=1}^{K} f_{ij}^2 p_{ij} = [(1-\Delta)X_t\{X_t/(X-1)-1\} + (1-p_{3\bullet})\alpha_t^*(1-\alpha_t^*) + \left(\alpha_t^*\right)^2]/(1-p_{3\bullet})^2$ and, through expression (C1), we obtain the following variance that yields the second expression (16)

$$V\left(\hat{\alpha}_t^*\right) = \frac{1}{n\left(1-p_{3\bullet}\right)^2}\left[\left(1-\Delta\right)X_t\left\{\frac{X_t}{X-1} - 1\right\} + \left(1-p_{3\bullet}\right)\alpha_t^*\left(1-\alpha_t^*\right)\right].$$

If $\hat{f} = \hat{\Delta}^*$, then $f = \Delta^* = (\Delta - \alpha_3)/(1-p_{3\bullet})$ and $f_{ij} = (\partial f/\partial \Delta)\times(\partial \Delta/\partial p_{ij}) + (\partial f/\partial \alpha_3)\times(\partial \alpha_3/\partial p_{ij}) + (\partial f/\partial p_{3\bullet})(\partial p_{3\bullet}/\partial p_{ij}) = [\Delta'_{ij} - \Delta'_{3(ij)} + \Delta^* \delta_{i3}]/(1-p_{3\bullet})$, where $\Delta'_{ij}$ and $\Delta'_{3(ij)}$ are given by expressions (C4) and (C6), respectively, for $R$=2 and $t$=3. By performing the operations $\sum_{i=1}^{K}\sum_{j=1}^{K} f_{ij} p_{ij} = -(1-\Delta^*)/(1-p_{3\bullet})$, $\sum_{i=1}^{K}\sum_{j=1}^{K} f_{ij}^2 p_{ij} = [\{B(1-X_3)(X-X_3)/(X-1)\} + (1-\Delta^*)(1-\Delta^* p_{3\bullet})]/(1-p_{3\bullet})^2$ and, through expression (C1), we obtain the following variance that yields expression (17)

$$\mathrm{V}\left(\hat{\Delta}^*\right) = \frac{1}{n\left(1-p_{3\bullet}\right)^2}\left[\left(1-\Delta\right)\left(1-X_3\right)\frac{X-X_3}{X-1} + \left(1-p_{3\bullet}\right)\Delta^*\left(1-\Delta^*\right)\right].$$

## REFERENCES

Agresti A (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research 1(2)*, 201-218.

Agresti A (2013). *Categorical Data Analysis* (3nd edn). Wiley: New Jersey. ISBN: 978-0-470-46363-5.

Agresti A and Lang JB (1993). Quasi-symmetric latent class models, with application to rater agreement. *Biometrics 49*, 131-139.

Agresti A, Ghosh A and Bini M (1995). Raking Kappa: Describing Potential Impact of Marginal Distributions on Measures of Agreement. *Biometrical Journal 37*, 811-820.

Asselineau J, Paye A, Bessède E, Perez P and Proust-Lima C (2018). Different latent class models were used and evaluated for assessing the accuracy of campylobacter diagnostic tests: overcoming imperfect reference standards? *Epidemiology and Infection 146*, 1556–1564

Ato M, López JJ and Benavente A (2011). A simulation study of rater agreement measures with 2×2 contingency tables. *Psicológica 32*, 385-402.

Banerjee M, Capozzoli M, McSweeney L and Sinha D (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics 27(1)*, 3-23.

Barlow W (2005). Modeling of categorical agreement. In: Armitage P, Colton T (eds), *The Encyclopedia of Biostatistics* (pp. 137-141). New York: Wiley.

Becker MP (1989). Using association models to analyse agreement data: two examples. *Statistics in Medicine 8*, 1199-1207.

Brenann RL and Prediger DJ (1981). Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement* 41, 687-699.

Cicchetti DV and Feinstein AR (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiolog 43*, 551-558

Cohen J (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*, 37-46.

Conger AJ (1980). Integration and generalization of kappas for multiple raters. *Psychological*

*Bulletin 88,* 322-328. DOI: 10.1037/0033-2909.88.2.322.

Conger AJ (2017). Kappa and Rater Accuracy: Paradigms and Parameters. *Educational and Psychological Measurement* 77(6), 1019-1047.

Darroch JN and McCloud PI (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics 28*, 371-388.

Dillon WR and Mulani N (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research 19*, 438-458.

Erdmann TP, De Mast J and Warrens MJ (2015). Some common errors of experimental design, interpretation and inference in agreement studies. *Statistical Methods in Medical Research 24 (6)*, 920-935.

Fleiss JL (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*, 378-382.

Fleiss JL, Levin B and Paik MC (2003). *Statistical methods for rates and proportions*. (3nd Ed.) John Wiley & Sons. ISBN: 978-0-471-52629-2.

Goodman LA (1991). Measures, models, and graphical displays in the analysis of cross-classified data (with discussion). *Journal of the American Statistical Association 86*, 1085-1111.

Guggenmoos-Holzmann I and Vonk, R (1998). Kappa-Like Indices of Observer Agreement Viewed From a Latent Class Perspective. *Statistics in Medicine 17*, 797-812.

Hubert L (1977). Kappa revisited. *Psychological Bulletin 48(2)*, 289-297.

Klauer KC and Batchelder WH (1996). Structural analysis of subjective categorical data. *Psychometrika 61(2)*, 199-240.

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

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

Lord FM and Novick MR (1968). *Statistical Theories of Mental Test Scores*. Menlo Park, Ca: Addison-Wesley.

Marasini D, Quatto P and Ripamonti E (2016). Assessing the inter-rater agreement for ordinal data through weighted indexes. *Statistical Methods in Medical Research 25(6)*, 2611-2633.

Martín Andrés A and Álvarez Hernández M (2019). Hubert's multi-rater kappa revisited. To appear in *British Journal of Mathematical and Statistical Psychology*. DOI: 10.1111/bmsp.12167.

Martín Andrés A and Femia Marzo P (2004). Delta: a new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology 57 (1)*, 1-19.

Martín Andrés A and Femia Marzo P (2005). Chance-corrected measures of reliability and validity in K×K tables. *Statistical Methods in Medical Research 14*, 473-492.

Martín Andrés A and Femia Marzo P (2008). Chance-corrected measures of reliability and validity in 2×2 tables. *Communications in Statistics-Theory and Methods 37*, 760-772. DOI: 10.1080/03610920701669884. Corrigendum in 39, 3009 (2010).

Martín Andrés A and Luna del Castillo JL (1989). Tests and intervals in multiple choice tests: A modification of the simplest classical model. *British Journal of Mathematical and Statistical Psychology 42*, 251-263.

Martín Andrés A and Luna del Castillo JL (1990). Multiple choice tests: Power, length and optimal number of choices per item. *British Journal of Mathematical and Statistical Psychology 43,* 57-71.

Nelson JC and Pepe MS (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research 9*, 475-496.

Qu Y, Tan M and Kutner MH (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics 52*, 797-810.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, J. Neyman (ed.), 321-333. Berkeley, California: University of California Press.

Shankar V and Bangdiwala SI (2014). Observer agreement paradoxes in 2x2 tables: comparison of agreement measures. *BMC Medical Research Methodology 14*, 100.

Scott WA (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opin Q, 19*, 321-325.

Schuster C and Smith DA (2005). Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. *Psychometrika 70(1)*, 135-146.

Stoyan D, Pommerening A, Hummel M and Kopp-Schneider A (2018). Multiple-rater kappas for binary data: Models and interpretation. *Biometrical Journal 60*, 381-394.

Tanner MA and Young MA (1985a). Modelling agreement among raters. *Journal of the American Statistical Association 80*, 175-180.

Tanner MA and Young MA (1985b). Modeling ordinal scale disagreement. *Psychological Bulletin 98*, 408-415.

Uebersax JS (1992). A review of modelling approaches for the analysis of observer agreement. *Investigative Radiology 27*, 738-743.

Uebersax JS and Grove WM (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine 9*, 559-572.

Vach W (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology 58*, 655-661.

Warrens MJ (2010). Inequalities between multi-rater kappas. *Adv Data Anal Classif 4*, 271-286.

**Table 1**

**Diagnosis of *n*=100 subjects by *R*=2 raters in *K*=3 categories**

**(a) Observed proportions ( $\bar{p}_{ij}$ )**

| *Rater 1* | *Rater 2* | | | *Totals* |
|---|---|---|---|---|
| | *Psychotic* | *Neurotic* | *Organic* | ( $\bar{p}_{i\bullet}$ ) |
| *Psychotic* | .75 | .01 | .04 | .80 |
| *Neurotic* | .05 | .04 | .01 | .10 |
| *Organic* | .00 | .00 | .10 | .10 |
| *Totals* ( $\bar{p}_{\bullet j}$ ) | .80 | .05 | .15 | 1 ( $\bar{p}_{\bullet\bullet}$ ) |

**(b) Data needed to make inferences with the *multi-rater delta* model. The data (raw) are obtained based on the data in Table 1(a)**

| *Categories* (*i*) | *Agreements* $\bar{p}_i$ | *Disagreements of rater r in category i* $\bar{d}_{ir}$ | | *Totals* $\bar{D}_i$ |
|---|---|---|---|---|
| | | *Rater=1* | *Rater=2* | |
| *Psychotic (1)* | .75 | .05 | .05 | .10 |
| *Neurotic (2)* | .04 | .06 | .01 | .07 |
| *Organic (3)* | .10 | .00 | .05 | .05 |
| *Totals* | $\bar{p}$ =.89 | $\bar{D}$=.11 | $\bar{D}$=.11 | 2 $\bar{D}$=.22 |

**(c) Estimate of the parameters and the measures of the degree of agreement of the *multi-rater delta* model for the data in Table 1(b). In the case of measures of degree of agreement, it also indicates their SEs**

| *Categories* (*i*) | *Parameters* $\hat{\alpha}_i$ | *Parameters* $\hat{\pi}_{ir}$ | | *Consistencies* $\left(\hat{\mathcal{S}}_i \pm SE\right)$ |
|---|---|---|---|---|
| | | *r=1* | *r=2* | *(degree of agreement in class i)* |
| *Psychotic (1)* | .5500 | .8000 | .8000 | .6875±.1442* |
| *Neurotic (2)* | .0375 | .2000 | .0400 | .5000±.2058* |
| *Organic (3)* | .1000 | .0000 | .1600 | .8000±.1085* |
| *Overall degree of agreement* ( $\hat{\Delta}$±*SE*) | | | | .6875±.1099* |

* At least one estimation of $\hat{\pi}_{ir}$ is zero. Therefore, the values of *SE* have been obtained adding +.5 to all the observations $x_{ij}$.

**Table 2**

**Cognitive response cross-classification of $n$=164 subjects by $R$=3 raters in $K$=3 categories (Dillon and Mulani, 1984, p.449)**

**(a) Absolute frequencies $x_{i_1 i_2 i_3}$. Observed proportions are $\bar{p}_{i_1 i_2 i_3} = x_{i_1 i_2 i_3} / n$**

| *Rater 3* | | *1* | | | *2* | | | *3* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Rater 2* | | *1* | *2* | *3* | *1* | *2* | *3* | *1* | *2* | *3* |
| | *1* | 56 | 1 | 0 | 5 | 3 | 0 | 0 | 0 | 1 |
| *Rater 1* | *2* | 12 | 2 | 1 | 14 | 20 | 4 | 0 | 4 | 2 |
| | *3* | 1 | 1 | 0 | 2 | 1 | 7 | 2 | 1 | 24 |

**(b) Data needed to apply the *multi-rater delta* model, which are obtained from Table 2(a)**

| *Categories* ($i$) | *Agreements* $n\,\bar{p}_i$ | *Disagreements of rater r in category i* $n\,\bar{d}_{ir}$ | | | *Total disagreements* $n\bar{D}_i$ |
|---|---|---|---|---|---|
| | | *Rater=1* | *Rater=2* | *Rater=3* | |
| *1* | 56 | 10 | 36 | 18 | 64 |
| *2* | 20 | 39 | 13 | 36 | 88 |
| *3* | 24 | 15 | 15 | 10 | 40 |
| *Totals* | $n\,\bar{p}$=100 | $n\,\bar{D}$=64 | $n\,\bar{D}$=64 | $n\,\bar{D}$=64 | $nR\,\bar{D}$=192 |

**(c) Estimate of the parameters and the measures of the degree of agreement of the *multi-rater delta* model for the data in Table 2(b). In the case of measures of degree of agreement, it also indicates their SEs**

| *Categories (i)* | *Parameters* $\hat{\alpha}_i$ | *Parameters* $\hat{\pi}_{ir}$ | | | *Consistencies* $\left(\hat{\mathcal{S}}_i \pm SE\right)$ *(degree of agreement in class i)* |
|---|---|---|---|---|---|
| | | *r=1* | *r=2* | *r=3* | |
| *1* | .3320 | .1564 | .5084 | .2647 | .7040±.0460 |
| *2* | .0741 | .6343 | .2823 | .5937 | .2462±.1011 |
| *3* | .1435 | .2093 | .2093 | .1416 | .6306±.0668 |
| *Overall degree of agreement* ($\hat{\Delta}$±*SE*) | | | | | .5496±.0462 |

**(d) Degree of overall agreement and per category when the *multi-rater delta* model is applied to the data in Table 2(a) collapsed on only two of the raters. As a reference, the same measurements are also provided when three raters are considered**

| *Categories (i)* | *Consistencies* | | | |
|---|---|---|---|---|
| | *Raters 1,2 and 3* | *Raters 1 and 2* | *Raters 1 and 3* | *Raters 2 and 3* |
| *1* | .704 | .706 | .788 | .810 |
| *2* | .246 | .160 | −.474 | −.570 |
| *3* | .631 | .764 | .714 | .709 |
| *Overall* | .550 | .567 | .329 | .413 |

**(e) Data needed to estimate Hubert's *kappa* (*R*-wise), which are obtained from Table 2(b)**

| *Categories* ($i$) | *Agreements* $n\,\overline{p}_i$ | *Number of responses i from rater r* $n\,\overline{t}_{ir} = n\left(\overline{p}_i + \overline{d}_{ir}\right)$ | | |
|---|---|---|---|---|
| | | *Rater=1* | *Rater=2* | *Rater=3* |
| *1* | 56 | 66 | 92 | 74 |
| *2* | 20 | 59 | 33 | 56 |
| *3* | 24 | 39 | 39 | 34 |
| *Totals* | $n\,\overline{p}=100$ | $n$=164 | $n$=164 | $n$=164 |

**(f) Data needed to estimate Fleiss' *kappa*, which are obtained from Table 2(a)**

| *Categories* ($i$) | *Number of subjects* $s_{i\omega}$ *in which* $R_{si} = \omega$ *raters respond i (the value* $\omega$*=0 has no interest)* | | | *Total* $R_{\bullet i}$ *of responses i (all raters)* |
|---|---|---|---|---|
| | $\omega$=1 | $\omega$=2 | $\omega$=3=$R$ | $(\Sigma_\omega \omega s_{i\omega})$ |
| *1* | 26 | 19 | 56 | 232=$R_{\bullet 1}$ |
| *2* | 32 | 28 | 20 | 148=$R_{\bullet 2}$ |
| *3* | 14 | 13 | 24 | 112=$R_{\bullet 3}$ |
| *Totals* ($s_{\bullet\omega}$) | 72 | 60 | 100 | 492=$\sum_{i=1}^{3} R_{\bullet i} = \sum_{\omega=1}^{3} \omega s_{\bullet\omega}$ |

**(g) Summary of the different measures of degree of agreement for the data in Table 2(a)**

| | |
|---|---|
| *Raw* | .610 |
| *Multi-rater delta* | .550 |
| *Hubert's kappa* (*R-wise*) | .547 |
| *Hubert's kappa* (*pairwise*) | .581 |
| *Fleiss' kappa* | .578 |

**Table 3**
**Modification of the data in Table 2 to obtain unbalanced marginal distributions (*n*=164)**

**(a) Absolute frequencies $x_{i_1 i_2 i_3}$. Observed proportions are $\bar{p}_{i_1 i_2 i_3} = x_{i_1 i_2 i_3} / n$**

| *Rater 3* | | 1 | | | 2 | | | 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Rater 2* | | *1* | *2* | *3* | *1* | *2* | *3* | *1* | *2* | *3* |
| | *1* | 108 | 1 | 0 | 2 | 3 | 0 | 0 | 0 | 1 |
| *Rater 1* | *2* | 2 | 2 | 1 | 4 | 10 | 4 | 0 | 4 | 0 |
| | *3* | 2 | 1 | 0 | 7 | 1 | 2 | 4 | 1 | 4 |

**(b) Data needed to apply the *multi-rater delta* model, which are obtained from Table 3(a)**

| *Categories* (*i*) | *Agreements* $n\,\bar{p}_i$ | *Disagreements of rater r in category i* $n\,\bar{d}_{ir}$ | | | *Total disagreements* $n\bar{D}_i$ |
|---|---|---|---|---|---|
| | | *Rater=1* | *Rater=2* | *Rater=3* | |
| *1* | 108 | 7 | 21 | 9 | 37 |
| *2* | 10 | 17 | 13 | 23 | 53 |
| *3* | 4 | 18 | 8 | 10 | 36 |
| *Totals* | $n\,\bar{p}=122$ | $n\,\bar{D}=42$ | $n\,\bar{D}=42$ | $n\,\bar{D}=42$ | $nR\,\bar{D}=126$ |

**(c) Data needed to estimate Hubert's *kappa* (*R*-wise), which are obtained from Table 3(b)**

| *Categories* (*i*) | *Agreements* $n\,\bar{p}_i$ | *Number of responses i from rater r* $n\,\bar{t}_{ir}=n\left(\bar{p}_i+\bar{d}_{ir}\right)$ | | |
|---|---|---|---|---|
| | | *Rater=1* | *Rater=2* | *Rater=3* |
| *1* | 108 | 115 | 129 | 117 |
| *2* | 10 | 27 | 23 | 33 |
| *3* | 4 | 22 | 12 | 14 |
| *Totals* | $n\,\bar{p}=122$ | *n*=164 | *n*=164 | *n*=164 |

**(d) Data needed to estimate Fleiss' *kappa*, which are obtained from Table 3(a)**

| *Categories* (*i*) | *Number of subjects $s_{i\omega}$ in which $R_{si}=\omega$ raters respond i (the value ω=0 has no interest)* | | | *Total $R_{\bullet i}$ of responses i (all raters)* ($\Sigma_\omega \omega s_{i\omega}$) |
|---|---|---|---|---|
| | ω=1 | ω=2 | ω=3=*R* | |
| *1* | 23 | 7 | 108 | $361=R_{\bullet 1}$ |
| *2* | 17 | 18 | 10 | $83=R_{\bullet 2}$ |
| *3* | 20 | 8 | 4 | $48=R_{\bullet 3}$ |
| *Totals* ($s_{\bullet\omega}$) | 60 | 33 | 122 | $492=\Sigma_i R_{\bullet i}=\Sigma_\omega \omega s_{\bullet\omega}$ |

**(e) Summary of the different measures of degree of agreement for the data in Table 3(a). Unbalanced marginals do not affect the *multi-rater delta* coefficient but do affect the *kappa* coefficient**

| | |
|---|---|
| *Raw* | .7439 |
| *Multi-rater delta* | .7075 |
| *Hubert's kappa* (*R-wise*) | .5739 |
| *Hubert's kappa* (*pairwise*) | .5553 |
| *Fleiss' kappa* | .5538 |