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

A. Martín Andrés and M. Álvarez Hernández

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 $kappa$ ($R \geq 2$) coefficients, which have several limitations. In 2004, the $delta$ coefficient was defined for the case of $R = 2$, which did not have the limitations of Cohen’s $kappa$ coefficient. This article extends the coefficient $delta$ from $R = 2$ raters to $R \geq 2$ (coefficient multi-rater $delta$), demonstrating that it can be expressed in the kappa format and has the same advantages as the coefficient $delta$ with regard to the type $kappa$ classic coefficients: (i) it refers to the proportion of replies that are concordant not by chance; (ii) allow to obtain a parameter that faithfully measures the degree of agreement in each category; and (iii) it is not affected by the marginal imbalance.

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 [1–5]. 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 ($\kappa_C$) [6]. Although $Kappa$ is a very popular and easily calculated measure of agreement, it has several limitations [7–12]. The two most relevant limitations are its dependence on marginal distributions and the difficulty in measuring the degree of agreement for each category. Martín Andrés & Femia Marzo [11,13,14] proposed a response model ($delta$ model) that led to the measure of agreement $delta$ ($\Delta$). Because it does not have the limitations of $\kappa_C$ [15–17], 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

CONTACT A. Martín Andrés aamartina@ugr.es Bioestadística, Facultad de Medicina, Universidad de Granada, 18071, C8-01, Granada, Spain

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2021.2013485

© 2021 Informa UK Limited, trading as Taylor & Francis Group
Table 1. Diagnosis of \( n = 100 \) subjects by \( R = 2 \) raters in \( K = 3 \) categories [18].

(a) Observed proportions (\( \hat{p}_{ij} \))

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

(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).

Disagreements of rater \( r \) in category \( i \): \( \hat{d}_{ir} \)

| Categories (i) | Agreements \( \hat{p}_i \) | Rater = 1 | Rater = 2 | Totals \( \hat{D}_i \) |
|----------------|---------------------------|-----------|-----------|-----------------|
| Psychotic (1)  | .75                       | .05       | .05       | .10             |
| Neurotic (2)   | .04                       | .06       | .01       | .07             |
| Organic (3)    | .10                       | .00       | .05       | .05             |
| Totals         | \( \hat{p} = .89 \)       | \( \hat{D} = .11 \) | \( \hat{D} = .11 \) | \( 2\hat{D} = .22 \) |

(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 \( \hat{\alpha}_{ir} \) | Parameters \( \hat{\pi}_{ir} \) | Consistencies (\( \hat{S}_i \pm SE \)) (degree of agreement in class \( i \)) |
|----------------|-----------------------------------|---------------------------------|-----------------------------------------------|
| Psychotic (1)  | .5500                            | \( \hat{\alpha}_1 \)           | \( \hat{\pi}_1 \)                             | \( \hat{S}_1 \pm .6875 \)                  |
| Neurotic (2)   | .0375                            | \( \hat{\alpha}_2 \)           | \( \hat{\pi}_2 \)                             | \( \hat{S}_2 \pm .5000 \)                  |
| Organic (3)    | .1000                            | \( \hat{\alpha}_3 \)           | \( \hat{\pi}_3 \)                             | \( \hat{S}_3 \pm .8000 \)                  |
| Overall degree of agreement (\( \hat{\Delta} \pm SE \)) | \( \hat{\Delta} \)         | \( \hat{\Delta} \)             | \( \hat{\Delta} \)                            | \( \hat{\Delta} \pm .6875 \)                |

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

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 first aim of this article is redefine the model for the case in which neither rater is a gold-standard. In addition, this extension will identify some minor errors committed by Martín Andrés & Femia Marzo when estimating the parameters of their delta model. A traditional example is Table 1(a), which comes from the classic example of Fleiss et al. [18] 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 [1,4,19–21]. Although Fleiss’ kappa (\( \kappa_F \)) is the most popular measure, it is not an extension of \( \kappa_C \) since the value of \( \kappa_F \) is not equal to \( \kappa_C \) when \( R = 2 \) [20]; however, it is an extension of Scott’s pi (\( \pi \)) coefficient [22] (see the quotations from [4] and [21]). Conversely, the generalizations of Hubert’s kappa [19] and Conger’s kappa [20] 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 [21,23], such as their dependence on the marginal imbalance. The second objective of this article is to extend the delta model from \( R = 2 \) to \( R \geq 2 \) (multi-rater delta model). An example for \( R = K = 3 \) is listed in Table 2(a), from Dillon and Mulani [24]; this example is cited by Schuster and Smith [5] 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.
Table 2. Cognitive response cross-classification of $n = 164$ subjects by $R = 3$ raters in $K = 3$ categories [24, p.449].

(a) Absolute frequencies $x_{i123}$. Observed proportions are $\hat{p}_{i123} = x_{i123}/n$

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

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

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

(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 $\hat{\pi}_i$ | Consistencies $(\hat{\alpha}_i \pm SE)$ (degree of agreement in class $i$) |
|-----------------|---------------------------|-------------------------------------------------------------------|
| $1$             | $.3320$                   | $.1564$                             | $.5084$                           | $.2647$                           | $.7040 \pm .0460$ |
| $2$             | $.0741$                   | $.6343$                             | $.2823$                           | $.5937$                           | $.2462 \pm .1011$ |
| $3$             | $.1435$                   | $.2093$                             | $.2093$                           | $.1416$                           | $.6306 \pm .0668$ |
| Overall degree of agreement $(\hat{\Delta} \pm SE)$ | | | | | $.5496 \pm .0462$ |

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

| Categories $(i)$ | Agreements $n \hat{p}_i$ | Number of responses $i$ from rater $r$ [ $\omega = n$ ] |
|-----------------|--------------------------|-------------------------------------------------|
| $1$             | $56$                     | $66$                                           | $92$                             | $74$ |
| $2$             | $20$                     | $59$                                           | $33$                             | $56$ |
| $3$             | $24$                     | $39$                                           | $39$                             | $34$ |
| Totals          | $n \hat{p} = 100$        | $n = 164$                                      | $n = 164$                        | $n = 164$ |

(e) Data needed to estimate Fleiss’ kappa, which are obtained from Table 2(a)

| Categories $(i)$ | $\omega = 1$ | $\omega = 2$ | $\omega = 3$ | $R$ | Total $R_{\omega i}$ of responses $i$ (all raters) $(\Sigma_{\omega = 1}^{R} \omega S_{\omega i})$ |
|-----------------|---------------|---------------|---------------|-----|-------------------------------------------------|
| $1$             | $26$          | $19$          | $56$          | $232 = R_{\ast 1}$ |
| $2$             | $32$          | $28$          | $56$          | $148 = R_{\ast 2}$ |
| $3$             | $14$          | $13$          | $24$          | $112 = R_{\ast 3}$ |
| Totals $(S_{\omega i})$ | $72$           | $60$          | $100$         | $492 = \sum_{i=1}^{R} R_{\omega i} = \sum_{\omega=1}^{R} \omega S_{\omega i}$ |

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

- Raw: .6098
- Multi-rater delta: .5496
- Hubert’s kappa (R-wise): .5471
- Hubert’s kappa (pairwise): .5809
- Fleiss’ kappa: .5777
Additionally, 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. These expressions will be extended to the case \( R \geq 2 \).

Finally, the current delta model should be framed within the scope of the known models. The analysis of the existing nominal agreement between several raters can be approached from various perspectives [25–28]. The most convenient option is usually the latent class model, which considers the existence of unobserved or latent variables [24, 29, 30]. 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. 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 [13], which in turn is derived from the model by Martín Andrés and Luna del Castillo [31] for multiple choice tests (an extension of the classic model by Lord and Novick [32]). And precisely the models of Martín Andrés and Luna del Castillo [31] and the latent classes of Klauer and Batchelder [30] are formally the same.

This article is organized as follows: in Section 2, the classic models kappa and delta are introduced, as well as the new 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 several examples and finally, in Section 6, the conclusions are set out.

2. Models for two raters

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 \( i (i = 1, 2, \ldots, K) \) and rater 2 as type \( j (j = 1, 2, \ldots, K) \), which generates a table of absolute frequencies \( x_{ij} \) and a table of relative frequencies (observed cell proportions) \( \hat{p}_{ij} = x_{ij}/n \), with \( \sum_{i=1}^{K} \sum_{j=1}^{K} x_{ij} = n \) and \( \sum_{i=1}^{K} \sum_{j=1}^{K} \hat{p}_{ij} = 1 \). The notation for the observed totals of row (\( x_{i.} \) and \( \hat{p}_{i.} \)), column (\( x_{.j} \) and \( \hat{p}_{.j} \)) or overall (\( x_{..} = n \) and \( \hat{p}_{..} = 1 \)) 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 \( \{p_{ij}\}; \{p_{i.}\} \) and \( \{p_{.j}\} \) will be the marginal distributions of row raters and column raters, respectively.

To analyse the previous problem, Cohen [6] defined the classic kappa coefficient \( \kappa_C = (I_o - I_e)/(1 - I_e) \) estimated by \( \hat{\kappa}_C = (\bar{I}_o - \bar{I}_e)/(1 - \bar{I}_e) \), where \( I_o = \sum_{i=1}^{K} \sum_{j=1}^{K} p_{ij} \) \( \bar{I}_o \) is the real (estimated) observed agreement index, and \( I_e = \sum_{i=1}^{K} \hat{p}_{i.} \hat{p}_{.i} \) \( \bar{I}_e \) 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 \( (1 - \bar{I}_e) \) has the two following effects on \( \hat{\kappa}_C \) [13]; 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 \( (1 - \bar{I}_e) \); 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’.
One way to correct these two limitations is by using the following *delta* model [13]:

\[ p_{ij}/p_{i*} = \delta_{ij} \Delta_i + (1 - \Delta_i) \pi_j, \]  

(1)

where \( \delta_{ij} \) allude to the Kronecker delta, \( 0 \leq \pi_j \leq 1, \sum_{j=1}^{K} \pi_j = 1 \) and \( \Delta_i \leq 1 \). This model is based on the model that Martín Andrés and Luna del Castillo [31] 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 [32]; the model of Martín Andrés and Luna del Castillo [31] is the same as the model of the expression (1) of Klauer and Batchelder [30], 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 as type \( i \) 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 \( p_{ii} = p_{i*} \Delta_i + p_{*i} (1 - \Delta_i) \pi_j \), the proportion of agreements in category \( i \) is the sum of the proportions of agreements that are not due to chance -the first summand \( p_{i*} \Delta_i \)- and the proportions of agreements due to chance -the second summand \( p_{*i} (1 - \Delta_i) \pi_j \). Hence, the total proportion of agreements not occurring by chance will be \( \Delta = \sum p_{i*} \Delta_i \). The authors [11] 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 \( \hat{\kappa}_C = .6765 \) despite the imbalance of the two marginals, but in other cases the values can be very different (see the second paragraph of Section 5). However, the interpretation is now more direct: 68.75% of the responses are concordant beyond chance (as opposed to 89% of raw agreements).

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:

\[ p_{ij} = \delta_{ij} \alpha_i + (1 - \Delta) \pi_{ij} \]  

(2)

where \( \alpha_i \leq 1, \Delta = \sum \alpha_i \leq 1 \) (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). 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 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 two ‘raw’ parameters of interest. The first is the total proportion of agreements \( p = \sum_{i=1}^{K} p_{ii} \), where \( p_{ii} \) is the proportion of agreements in category \( i \). 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 \( p_{i*} \) and \( p_{*i} \) in category \( i \); for example, if \( p_{ii} = p_{i*} = p_{*i} = 0.01 \), then the degree of agreement in class \( i \) is maximal, but the proportion of agreements \( (p_{ii} = 0.01) \) is very small due to the fact that the marginal proportions are also very small. A suitable method for defining the degree of agreement in
category \( i \) (the second parameter) is through the consistency \( S_i = 2\pi_i/(p_{i\bullet} + p_{\bullet i}) \) of Martín Andrés & Femia Marzo [11], which is a parameter that is based on the 'proportion of specific agreement' of Fleiss et al. [18] and on the agreement index of Cichetti & Feinstein [33]. Note that \( S_i \) is the proportion of agreements obtained among all of the responses \( i \) of any one of the two raters. The two agreement parameters \( p \) 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 two previous raw parameters \( p_{ii} \) are replaced by \( \alpha_i \), then two parameters that are corrected by chance will be obtained (which is the current objective). In the new delta model for two raters, the consequence is that the two parameters of interest are the overall degree of agreement \( \Delta = \Sigma \alpha_i \) (that is, the total proportion of agreements that do not occur by chance), and the consistency \( S_i \) in category \( i \) (that is, the degree of agreement in category \( i \) that do not occur by chance),

\[
S_i = \frac{2\alpha_i}{p_{i\bullet} + p_{\bullet i}} = \frac{2\alpha_i}{2\alpha_i + (1 - \Delta)(\pi_{i1} + \pi_{i2})},
\]

where \( \alpha_i \) is the proportion of agreement in category \( i \) that do not occur by chance and the second equality is due to \( p_{i\bullet} = \Sigma_{j=1}^{K}p_{ij} = \alpha_i + (1 - \Delta)\pi_{i1} \) and \( p_{\bullet i} = \Sigma_{i=1}^{K}p_{ij} = \alpha_j + (1 - \Delta)\pi_{j2} \). In the new delta model, note that the possibility that \( S_i < 0, \alpha_i < 0 \) 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 and Femia Marzo [11] explained that \( S_i \) 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 \( S_i \); 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

Let there now be \( R \geq 2 \) raters who independently classify \( n \) subjects in \( K \) categories, producing a data matrix \( \{y_{sr}\} \), with \( n \) rows \( (s = 1, 2, \ldots, n) \), \( R \) columns \( (r = 1, 2, \ldots, R) \) and values \( y_{sr} = 1, 2, \ldots, K; \) 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 \( x_{i1i2...iR} = \#\{s|y_{s1} = i_1, \ldots, y_{sR} = i_R\} \) of dimension \( K^R \), where the symbol \# refers to 'cardinal', which incorporates the frequencies of all possible combinations of classifications carried out by the \( R \) raters. Therefore \( x_{i1i2...iR} \) is the number of subjects classified as type \( i_1 \) by rater 1, type \( i_2 \) by rater 2, \ldots, or type \( i_R \) by rater \( R \), and \( \{x_{i1i2...iR}\} \) are the observed values of a multinomial random variable of sample size \( n \) and probabilities \( \{p_{i1i2...iR}\} \). Let \( \tilde{p}_{i1i2...iR} = x_{i1i2...iR}/n \) be the observed cell proportion, \( \tilde{p}_i = \tilde{p}_{i1...i} = \#\{s|y_{s1} = \ldots = y_{sR} = i\}/n = x_{i...i}/n = x_i/n \) the observed proportion of agreements in category \( i \) and its average value \( \tilde{p}_i = \tilde{p}_{ii...ii}, \tilde{p} = \Sigma_{i=1}^{K} \tilde{p}_i \) the observed
depending on the definition of category and $\sum p_i$ is an estimator of the population coefficient of agreement of type $\kappa$.

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, $E(A_0)$ is the average value of $A_0$ on the assumption of independence among all the raters, then Hubert [19] indicated that the estimate of the degree of agreement of type $\kappa$ is given by

$$\hat{\kappa} = \frac{A_0 - E(A_0)}{\text{Max}(A_0) - E(A_0)} = 1 - \frac{\text{Max}(A_0) - A_0}{\text{Max}(A_0) - E(A_0)} = 1 - \frac{1 - \bar{I}_o}{1 - \bar{I}_c} = \frac{\bar{I}_o - \bar{I}_c}{1 - \bar{I}_c}$$

(4)

where $\bar{I}_o = A_0 / \text{Max}(A_0)$ and $\bar{I}_c = E(A_0) / \text{Max}(A_0)$. Hubert [19] makes the following interpretation ‘an agreement occurs if and only if all raters agree on the categorization of an object’. In this case, $\text{Max}(A_0) = n$, $A_o = n \bar{I}_o = n \sum_{i=1}^K \bar{p}_i = n \bar{p}$, $E(A_0) = n \bar{I}_c = n \sum_{i=1}^K \Pi_{r=1}^R \bar{t}_{ir}$ and Hubert’s $\kappa$ coefficients

$$\hat{\kappa}_{HR} = \frac{\bar{I}_o - \bar{I}_c}{1 - \bar{I}_c},$$

where $\bar{I}_o = \sum_{i=1}^K \bar{p}_i = \bar{p}$ and $\bar{I}_c = \sum_{i=1}^K \Pi_{r=1}^R \bar{t}_{ir}$, Martín Andrés and Álvarez Hernández [34] 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 [1] and Hubert [19] or a pairwise definition of agreement. An extension of the concept is Conger’s $g$-wise $\kappa$ [20], with $2 \leq g \leq R$, where $g = R$ or $g = 2$ yields the two previously mentioned Hubert definitions ($R$-wise and pairwise, respectively). However, $\kappa$ coefficients can vary from one author to another, depending on the definition of $I_c$. The most traditional definitions are those of the Fleiss $\kappa$ ($\kappa_F$) [1] and that of Hubert’s $\kappa$ ($\kappa_H$) [19] estimated by

$$\hat{\kappa}_F = 1 - \frac{n R^2 - \sum_{i=1}^n R_{si}}{n R (R - 1) \left(1 - \sum_{i=1}^K R_{ii}^2 / n\right)}$$

$$\hat{\kappa}_H = 1 - \frac{R^2 - \left(\sum_{i=1}^n R_{si}^2 / n\right)}{R (R - 1) - 2 \sum_{i=1}^K \sum_{r=1}^R \sum_{r'=r+1}^R \bar{t}_{ir} \bar{t}_{ir'}},$$

(6)

where $R_{si} = \#(r \mid y_{sr} = i) = 0, 1, \ldots, R$ is the number of raters that classify subject $s$ in category $i$, $R_i = R_{si} / n R$ is the total proportion of responses $i$ (any rater) and $R_{si} = \sum_{r=1}^R R_{sr}$. In the case of $\hat{\kappa}_H$, the expression derives from the fact that $A_0 = \sum_{i=1}^n \sum_{r=1}^R R_{sr}(R_{si} - 1) / 2$, $\text{Max}(A_0) = n R (R - 1) / 2$ and $E(A_0) = n \sum_{i=1}^K \sum_{r=1}^R \sum_{r'=r+1}^R \bar{t}_{ir} \bar{t}_{ir'}$. Warrens [4] proved that $\hat{\kappa}_F \leq \hat{\kappa}_H$ and that if all the raters classify all the subjects, then $\hat{\kappa}_H$ is more appropriate. It can be seen that when $R = 2$ then $\hat{\kappa}_{HR} = \hat{\kappa}_H = \hat{\kappa}_C$, but $\hat{\kappa}_F \neq \hat{\kappa}_C$. 
The extension of the multi-rater delta model for \( R = 2 \)-expressions (2)- to the case of many raters is immediate. Now

\[
p_{i_1i_2...i_R} = \delta_{i_1i_2...i_R} \alpha_i + (1 - \Delta) \prod_{r=1}^{R} \pi_{ir} \quad \text{with} \quad \Delta = \sum_{i=1}^{K} \alpha_i \quad \text{and} \quad 1 = \sum_{i_r=1}^{K} \pi_{i_r} \quad (7)
\]

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

\[
S_i = \frac{R \alpha_i}{\sum_{r=1}^{R} t_{ir}} = \frac{R \alpha_i}{R \alpha_i + (1 - \Delta) \sum_{r=1}^{R} \pi_{ir}} \quad (8)
\]

In the multi-rater delta model, note that the possibility that \( S_i < 0 \), \( \alpha_i < 0 \) or \( \Delta < 0 \) is also allowed because the degree of agreement can sometimes be negative.

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

\[
\Delta = \frac{I_{\alpha} - I_{\pi}}{1 - I_{\pi}} \quad \text{where} \quad I_{\pi} = \sum_{i=1}^{K} \prod_{r=1}^{R} \pi_{ir}, \quad (9)
\]

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 \( K > 2 \))

As can be seen in Appendix B, and because of the structure of the model, to make inferences with the multi-rater delta model, only some of the observed cell proportions \( \hat{p}_{i_1i_2...i_R} \) are needed: \( \hat{p}_i \) (the observed proportions of agreements in category \( i \)) and \( \hat{d}_{ir} = \hat{t}_{ir} - \hat{p}_i \) (the observed proportions of disagreements by rater \( r \) in category \( i \)). Based on these proportions, the total proportion of agreements \( \hat{p} = \sum_{i=1}^{K} \hat{p}_i \), the total proportion of disagreements \( \hat{D} = \sum_{i=1}^{K} \hat{d}_{ir} \) (which is the same for all raters), and the total proportion of disagreements in category \( i \) \( \hat{D}_i = \sum_{r=1}^{R} \hat{d}_{ir} \) are determined. Based on the definitions, \( \hat{t}_{ir} = \hat{p}_i + \hat{d}_{ir} \), \( 1 = \hat{p} + \hat{D} \), \( \hat{D} = \sum_{i=1}^{K} \hat{D}_i/R \) and \( \hat{D}_i \leq (R - 1)\hat{D} \). In addition, it can be seen that the values \( R_i \) of \( \hat{k}_{ir} \) are given by \( R_i = \hat{p}_i + \hat{D}_i/R \).

Once these proportions are known, Appendix B shows that the estimators of the maximum likelihood (\( \hat{\pi}_{ir}, \hat{\alpha}_i, \hat{\Delta} \) and \( \hat{S}_i \)) of the various parameters of interest (\( \pi_{ir}, \alpha_i, \Delta \) and \( S_i \)) are as follows:

\[
\begin{align*}
\hat{\pi}_{ir} &= \frac{\lambda_i + \hat{d}_{ir}}{B}, \quad \hat{\alpha}_i = \frac{\hat{p}_i - \lambda_i}{B} = 1 - B \quad \text{and} \quad \hat{\Delta} = \frac{R \hat{\alpha}_i}{\sum_{r=1}^{R} t_{ir}} \quad \text{with} \quad \hat{D} = \frac{R \hat{\alpha}_i}{\sum_{r=1}^{R} t_{ir}} \\
&= \frac{R \hat{\alpha}_i}{R \hat{p}_i + \hat{D}_i} = \frac{R \hat{\alpha}_i}{R \hat{p}_i + (1 - \Delta) \sum_{r=1}^{R} \hat{\pi}_{ir}} \quad (10)
\end{align*}
\]
where \( B = 1 - \Delta \geq 0 \) and \( \lambda_i = B \sum_{r=1}^{R} \pi_{ir} \geq 0 \) are the solutions of expressions

\[
B^{R-1} = \frac{\prod_{r=1}^{R} (\lambda_i + \bar{d}_{ir})}{\lambda_i} (\forall i | \bar{d}_{ir} \neq 0, \forall r) \quad \text{under the condition } g(B)
\]

\[
= \sum_{i=1}^{K} \lambda_i - B + \bar{D} = 0
\]

with the exception that \( \lambda_i = 0 \) when \( \bar{d}_{ir} = 0 \) for some rater \( r \). In addition, the Supplementary Material explicitly sets out how to proceed in order to determine the values of \( B \) and \( \lambda_i \), with details on how to act in the extreme cases of \( B = 0 \) and \( B = \infty \). In the special case of independence among raters, that is, \( \bar{p}_i = \prod_{r=1}^{R} \bar{t}_{ir} = \prod_{r=1}^{R} (\bar{p}_i + \bar{d}_{ir}) \), then \( \lambda_i = \bar{p}_i \) and \( B = 1 \) are the solutions of expressions (11); thus, \( \hat{\alpha}_i = \hat{\Delta} = \hat{k}_{HR} = 0 \). When \( \bar{I}_0 = \sum_{i=1}^{K} \bar{p}_i = 1 \), in which case \( \hat{k}_{HR} = 1 \), then \( \bar{d}_{ir} = 0 \) \( (\forall i, r) \), \( \lambda_i = 0 \) \( (\forall i) \), \( g(B) = 0 \) implies that \( B = \bar{D} = 0 \) and \( \hat{\alpha}_i = \hat{\Delta} = \hat{k}_{HR} = 1 \). In general, \( \hat{\Delta} \) and \( \hat{k}_{HR} \) take similar values, except when the marginals are very unbalanced, in which case \( \hat{\Delta} \) is usually superior to \( \hat{k}_{HR} \) (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 \( n\bar{p}_{i_1i_2...i_R} \) and \( n\hat{p}_{i_1i_2...i_R} = n\delta_{i_1i_2...i_R} \hat{\alpha}_i + n(1 - \hat{\Delta}) \prod_{r=1}^{R} \bar{p}_{i_r} \), respectively, then the type I \( \alpha \) error test consists of comparing

\[
\chi^2_{\text{exp}} = n \left[ \sum_{i_1=1}^{K} \ldots \sum_{i_R=1}^{K} \frac{(\bar{p}_{i_1...i_R} - \hat{p}_{i_1...i_R})^2}{\hat{p}_{i_1...i_R}} \right]
\]

\[
= n \left[ \frac{1 - \delta_{i_1...i_R}}{1 - \hat{\Delta}} \sum_{i_1=1}^{K} \ldots \sum_{i_R=1}^{K} \frac{\bar{p}_{i_1...i_R}^2}{\prod_{r=1}^{R} \bar{p}_{i_r}} - \bar{D} \right]
\]

vs the \((1-\alpha)\)-percentile \( \chi^2_{\alpha, df} \), 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 \( (\bar{p}_i = \hat{p}_i) \) and disagreements \( (\bar{d}_{ir} = \hat{d}_{ir}) \) are equal, which implies that the observed and expected marginal distributions are also equivalent \( (\bar{t}_{ir} = \hat{t}_{ir}) \). 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 \( R(K-1) \) probabilities \( \pi_{ir} \) estimated. Applying the classic validity rule, the previous goodness-of-fit test will be valid if none of the expected quantity \( n\hat{p}_{i_1i_2...i_R} \) is lower than 1 and no more than 20% of them are lower than or equal to 5. When the overall degree of agreement is high and the sample size is small, many of the observed and expected frequencies will be small and the previous test will not be reliable (at least in its significant results).

The variances of \( \hat{\Delta} \), \( \hat{\alpha}_i \) and \( \hat{S}_i \) must be obtained to make inferences about the measures of agreement \( \Delta \), \( \alpha_i \) and \( S_i \). Appendix C shows that

\[
\hat{V}(\hat{\Delta}) = \frac{1 - \hat{\Delta}}{n} \left\{ \hat{\Delta} + \frac{\hat{X}}{(R-1)\hat{X} - 1} \right\}
\]

(13)
\[
\hat{V}(\hat{\alpha}_i) = \frac{1}{n} \left[ \hat{\alpha}_i(1 - \hat{\alpha}_i) + (1 - \hat{\Delta})\hat{X}_i \left\{ \frac{(R - 1)\hat{X}_i}{(R - 1)\bar{X} - 1} - 1 \right\} \right] \quad \text{and} \quad \hat{V}(\hat{S}_i) = \frac{R^2}{nN_i^2} \left[ n\hat{V}(\hat{\alpha}_i) - \hat{\alpha}_i(1 - \hat{\alpha}_i) + \hat{\alpha}_i(1 - \hat{S}_i) \left\{ 1 - \frac{R - 1}{R} \hat{S}_i \right\} + \frac{(1 - \hat{\Delta})\hat{S}_i^2}{R^2} \left\{ \left( \sum_{r=1}^{R} \hat{\pi}_{ir} \right)^2 - \left( \sum_{r=1}^{R} \hat{\pi}_{ir}^2 \right) \right\} \right]
\]

where \( \hat{X}_i = \left[ \sum_{r=1}^{R} \hat{\pi}_{ir}^{-1} - (\prod_{r=1}^{R} \hat{\pi}_{ir})^{-1} \right]^{-1} \) 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 \( \tilde{d}_{ir} = 0 \) or \( B = \infty \) 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 \( x_{i12...ir} \) increased by .5; thus, the new sample size is \( n + KR/2 \), and the new proportions observed are \( \tilde{p}_{i12...ir} = (x_{i12...ir} + 0.5)/(n + KR/2) \).

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 [11,13]. 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. Special 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 \( (\alpha_1, \alpha_2, \pi_{11} \text{ and } \pi_{12}) \) than free cells to take values (of which there are only three). In this case, the following solution by Martín Andrés & Femia Marzo [13] can be adopted; it has been proved to provide coherent results [11,13–17]. 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 \( 3 \times 3 \) table by .5, estimate the parameters as executed in the previous section, and redefine the measures of agreement without considering the third dummy category. Let \( \tilde{p}_i \) and \( \tilde{d}_{ir} \) 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 \( 3 \times 3 \) table. The measures of agreement for the original \( 2 \times 2 \) table are defined as \( \alpha^*_i = \alpha_i/(p_{1\bullet} + p_{2\bullet}), \Delta^* = \alpha^*_1 + \alpha^*_2 \) and \( S^*_i = 2\alpha_i/(p_{1\bullet} + p_{2\bullet}) \), for \( i = 1 \) and \( 2 \); their estimates and estimated variances (see Appendix D) are

\[
\hat{\alpha}_i^* = \frac{\hat{\alpha}_i}{1 - \hat{p}_{3\bullet}}, \hat{V}(\hat{\alpha}_i^*) = \frac{1}{n(1 - \hat{p}_{3\bullet})^2} \left\{ (1 - \hat{\Delta})\hat{X}_i \left\{ \frac{\hat{X}_i}{\hat{X} - 1} - 1 \right\} + (1 - \hat{p}_{3\bullet})\hat{\alpha}_i^*(1 - \hat{\alpha}_i^*) \right\}, \quad \hat{\Delta}^* = \hat{\alpha}_1^* + \hat{\alpha}_2^*, \hat{V}(\hat{\Delta}^*)
\]
\[ \hat{k}_i = \frac{2\hat{\alpha}_i}{N_i}, \hat{V}(\hat{k}_i^*) \]
\[ = \frac{4}{nN_i^2} \left[ (1 - \hat{\Delta})\hat{X}_i \left\{ \frac{\hat{X}_i}{\hat{X}_i - 1} - 1 \right\} + \hat{\alpha}_i \left\{ 1 - \frac{3\hat{\alpha}_i}{N_i} + \frac{2\hat{\alpha}_i^2}{N_i^2} + \frac{2\hat{\pi}_{i1}\hat{\pi}_{i2}(1 - \hat{\Delta})\hat{\alpha}_i}{N_i^2} \right\} \right], \]

where \((1 - \hat{p}_3) = \hat{p}_1 + \hat{p}_2 + \hat{d}_{11} + \hat{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 that \(\hat{k}_C = .6765\). With only two raters, \(\kappa_F\) is not usually calculated; its estimation \(\hat{k}_F = .6753\) 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 \(\hat{p}_1 = \hat{p}_{11} = .75, \hat{d}_{11} = \hat{p}_{12} + \hat{p}_{13} = .05, \) etc. Based on this table and the expression (11), the values \(B = .3125, \lambda_1 = .0500, \lambda_2 = .0048, \) and \(\lambda_3 = 0\) (which is directly derived from the fact that \(\hat{d}_{31} = 0\)) 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 \(\hat{k}_C\)) 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 \(\hat{S}_1 = .6875, \hat{S}_2 = .5000\) and \(\hat{S}_3 = .8000\); 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: \(\hat{k}_{C(1)} = .6875, \hat{k}_{C(2)} = .5000\) and \(\hat{k}_{C(3)} = .7727\). According to Martín Andrés & Femia Marzo [11] 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 \(\hat{d}_{31} = 0\) (which is due to the fact that \(\lambda_3 = 0\)). 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 \(\hat{k}_C = .6765\) 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 [11,13], 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 \(\hat{k}_C = .4792\) and \(\hat{\Delta} = .9200\), 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 \emph{multi-rater delta} model Table 2(b) must first be constructed; for example, \( n \hat{\pi}_1 = 56 \) and \( n \hat{\delta}_{11} = 1 + 0 + 5 + 3 + 0 + 0 + 0 + 1 = 10 \). Based on this table, and the expression (11), the values \( B = 0.4504, \lambda_1 = 0.0104, \lambda_2 = 0.0774 \) and \( \lambda_3 = 0.0030 \) 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 \( -\hat{k}_{HR(1)} = 63.62\%, \hat{k}_{HR(2)} = 42.70\% \) and \( \hat{k}_{HR(3)} = 60.81\% \), especially in the case of 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 \( \frac{100}{164} = 60.98\% \) for the total degree of agreement and \( 3 \times 56/232 = 72.41\%, 3 \times 20/148 = 40.54\%, \) and \( 3 \times 24/112 = 64.29\% \) for consistencies in categories 1, 2, and 3, respectively. The estimated total degree of agreement is \( 0.5496 \pm 0.0462 \), which indicates that the real total degree of agreement is \( \Delta \geq 0.5496 - 1.645 \times 0.0462 = 0.4736 \) 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 \emph{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 \( R > 2 \).

Determining the different \( \kappa \) measures of agreement requires summarizing the data of Table 2(a) in a different way. To calculate \( \hat{k}_{HR} \) one needs the data in Table 2(d), which are obtained from Table 2(b); based on this and the expression (5), \( \hat{k}_{HR} = .5471 \). To calculate \( \hat{k}_F \) one needs the data in Table 2(e), which are obtained from Table 2(a); for example, the number 26 in Table 2(e), which is the number of subjects classified in category 1 by only one of the three raters, is obtained by summing all the frequencies \( x_{113} \) 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(e),

\[
\sum_{s=1}^{164} \sum_{i=1}^{3} R_{si}^2 = 72 \times 1^2 + 60 \times 2^2 + 100 \times 3^2 = 1212, \quad \sum_{i=1}^{3} R_i^2 = (232^2 + 148^2 + 112^2)/(164 \times 3)^2 = .3647 \text{ and, through the expression (6), } \hat{k}_F = 1 - [164 \times 9 - 1212]/[164 \times 3 \times 2 \times (1 - .3647)] = 0.5777.
\]

Finally, to calculate \( \hat{k}_{H2} \) the data in Table 2(d,e) is needed; from the expression (5), \( \hat{k}_{H2} = .5809 \). Table 2(f) 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 \( \sum_{i=1}^{3} \hat{\pi}_i = (56 + 20 + 24)/164 = .6098 \).

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

(a) Absolute frequencies \(x_{ijkl} \). Observed proportions are \(\bar{p}_{ijkl} = x_{ijkl} / n\)

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

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

| Categories (i) | Disagreements of rater r in category i | n \(\bar{d}_r\) | Total disagreements n\(\bar{D}\) |
|----------------|----------------------------------------|----------------|------------------|
| 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 | n\(\bar{D}\) = 126 |

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

| Categories (i) | Number of responses i from rater r |
|----------------|-----------------------------------|
|                | n \(\bar{p}\) = n (\(\bar{p}_i + \bar{d}_r\)) |
| 1              | 108                                |
| 2              | 10                                 |
| 3              | 4                                  |
| Totals         | n = 164                            |

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

| Categories (i) | \(\omega = 1\) | \(\omega = 2\) | \(\omega = 3\) = R | Total \(R_{\omega}\) of responses i |
|----------------|----------------|----------------|-------------------|----------------------------------|
| 1              | 23             | 7              | 108               | 361                             |
| 2              | 17             | 18             | 10                | 83                              |
| 3              | 20             | 8              | 4                 | 48                              |
| Totals (s_{\omega}) | 60            | 33             | 122               | 492                             |

(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

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

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% and 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 \(\chi^2_{\text{exp}} = 19.83\) (df = 17) is not significant for \(\alpha = 5\%\), even though 9 (24) expected amounts less than one (less than or equal to five) exist. At the end of Appendix B, a theoretical study is carried out on the value of the difference \(\hat{\Delta} - \hat{k}_{HR}\).
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 [6] and, occasionally, the \( \delta \) coefficient [13]. The classic \( \delta \) coefficient has three advantages over the \( \kappa \) coefficient. The first advantage is that the \( \delta \) coefficient is not affected by imbalance in the marginal distributions of each rater. The second advantage is that the \( \delta \) model from which this coefficient is derived has no difficulty in evaluating the degree of agreement in each category [11]. The opposite occurs in the case of the \( \kappa \) coefficient [7–9]. Finally, the \( \delta \) coefficient is easier to interpret than the \( \kappa \) coefficient; both refer to the proportion of replies that are concordant not by chance, but population references are different: the first one is the total number of responses (\( n \)) and the second one is the maximum number possible of non-random agreements \( (1 - \bar{I}_e) \). In the first case, the coefficient is interpreted as ‘\( \Delta \times 100\% \) of the responses are concordant not by chance’; in the second case ‘the two raters agree on \( \kappa_C \times 100\% \) of the maximum number possible of non-random agreements’.

When many raters (\( R \geq 2 \)) exist, several \( \kappa \)-based coefficients can be employed, such as Fleiss’ \( \kappa \) [1], Hubert’s \( \kappa \) [19], and Conger’s \( \kappa \) [20]. However, Fleiss’ \( \kappa \) does not coincide with Cohen’s \( \kappa \) when \( R = 2 \). The limitations of all multi-rater \( \kappa \) coefficients are identical to the limitations of the Cohen’s \( \kappa \) coefficient [21,23]. In this article, the classic \( \delta \) 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 provides results that are compatible with the results of the classic \( \delta \) model, so it has the same 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 \( S_i \)). 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. Lastly, the interpretation of the multi-rater \( \delta \) coefficient is also simpler than the interpretation of multi-rater \( \kappa \) coefficients. A free programme (Multi_Rater_Delta.exe) may be downloaded to run this model at http://www.ugr.es/local/bioest/software (Section ‘Agreement Among raters’).

Acknowledgements

The authors wish to thank Dr Pedro Femia Marzo for his invaluable assistance in various aspects of this article.
Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

A. Martín Andrés http://orcid.org/0000-0002-2548-2638
M. Álvarez Hernández http://orcid.org/0000-0001-7524-3339

References

[1] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378–382.
[2] Landis JR, Koch GG. A review of statistical methods in the analysis of data arising from observer reliability studies (part I). Stat Neerl. 1975;29:101–123.
[3] Landis JR, Koch GG. A review of statistical methods in the analysis of data arising from observer reliability studies (part II). Stat Neerl. 1975;29:151–161.
[4] Warrens MJ. Inequalities between multi-rater kappas. Adv Data Anal Classif. 2010;4:271–286.
[5] Schuster C, Smith DA. Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. Psychometrika. 2005;70(1):135–146.
[6] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
[7] Brennann RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Meas. 1981;41:687–699.
[8] Agresti A, Ghosh A, Bini M. Raking kappa: describing potential impact of marginal distributions on measures of agreement. Biom J. 1995;37:811–820.
[9] Guggenmoos-Holzmann I, Vonk R. Kappa-like indices of observer agreement viewed from a latent class perspective. Stat Med. 1998;17:797–812.
[10] Nelson JC, Pepe MS. Statistical description of interrater variability in ordinal ratings. Stat Methods Med Res. 2000;9:475–496.
[11] Martin Andrés A, Femia Marzo P. Chance-corrected measures of reliability and validity in K×K tables. Stat Methods Med Res. 2005;14:473–492.
[12] Erdmann TP, De Mast J, Warrens MJ. Some common errors of experimental design, interpretation and inference in agreement studies. Stat Methods Med Res. 2015;24(6):920–935.
[13] Martin Andrés A, Femia Marzo P. Delta: a new measure of agreement between two raters. Br J Math Stat Psychol. 2004;57(1):1–19.
[14] Martin Andrés A, Femia Marzo P. Chance-corrected measures of reliability and validity in 2×2 tables. Commun Stat – Theory and Methods. 2008;37:760–772. Corrigendum in 39, 3009 (2010). doi:10.1080/03610920701669884.
[15] Ato M, López JJ, Benavente A. A simulation study of rater agreement measures with 2×2 contingency tables. Psicologica. 2011;32:385–402.
[16] Shankar V, Bangdiwala SI. Observer agreement paradoxes in 2×2 tables: comparison of agreement measures. BMC Med Res Methodol. 2014;14:100.
[17] Giammarino M, Mattiello S, Battini M, et al. Evaluation of inter-observer reliability of animal welfare indicators: which is the best index to use? Animals. 2021;11:1445.
[18] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. New Jersey: John Wiley & Sons; 2003. ISBN: 978-0-471-52629-2.
[19] Hubert L. Kappa revisited. Psychol Bull. 1977;48(2):289–297.
[20] Conger AJ. Integration and generalization of kappas for multiple raters. Psychol Bull. 1980;88:322–328. doi:10.1037/0033-2909.88.2.322.
[21] Marasini D, Quatto P, Ripamonti E. Assessing the inter-rater agreement for ordinal data through weighted indexes. Stat Methods Med Res. 2016;25(6):2611–2633.
[22] Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opin Q. 1955;19:321–325.
[23] Conger AJ. Kappa and rater accuracy: paradigms and parameters. Educ Psychol Meas. 2017;77(6):1019–1047.

[24] Dillon WR, Mulani N. A probabilistic latent class model for assessing inter-judge reliability. Multivariate Behav Res. 1984;19:438–458.

[25] Agresti A. Modelling patterns of agreement and disagreement. Stat Methods Med Res. 1992;1(2):201–218.

[26] Uebersax JS. A review of modelling approaches for the analysis of observer agreement. Invest Radiol. 1992;27:738–743.

[27] Banerjee M, Capozzoli M, McSweeney L, et al. Beyond kappa: a review of intrarater agreement. Can J Stat. 1999;27(1):3–23.

[28] Barlow W. Modeling of categorical agreement. In: Armitage P, Colton T, editors. The encyclopedia of biostatistics. New York: Wiley; 2005. p. 137–141.

[29] Lord FM, Novick MR. Statistical theories of mental test scores. Menlo Park (CA): Addison-Wesley; 1968.

[30] Klauer KC, Batchelder WH. Structural analysis of subjective categorical data. Psychometrika. 1996;61(2):199–240.

[31] Agresti A. Categorical data analysis. 3rd ed. New Jersey: John Wiley & Sons; 2013. ISBN: 978-0-470-46363-5.

[32] Barlow W. Modeling of categorical agreement. In: Armitage P, Colton T, editors. The encyclopedia of biostatistics. New York: Wiley; 2005. p. 137–141.

Appendices

Appendix 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 and Femia Marzo [11,13] defined the following three parameters for the case of any two raters

\[
\mathcal{A}_i = p_{i\bullet}\Delta_i, \quad \Delta = \sum \mathcal{A}_i \quad \text{and} \quad \mathcal{S}_i = 2p_{i\bullet}\Delta_i/(p_{i\bullet}+p_{\bullet i}) \tag{A1}
\]

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),

\[
p_{ij} = \delta_{ij}\alpha_i + (p_{i\bullet} - \alpha_i)\pi_j \quad \text{with} \quad \alpha_i = p_{i\bullet}\Delta_i = \mathcal{A}_i \tag{A2}
\]

If \( \Delta = \sum_{i=1}^{K} \alpha_i = \sum_{i=1}^{K} p_{i\bullet}\Delta_i \) from expression (A2), \( p_{ij} = \sum_{i=1}^{K} p_{ij} = p_{jj} + \sum_{i=1,i\neq j}^{K} p_{ij} = \alpha_j + (p_{i\bullet} - \alpha_i)\pi_j + \pi_j \sum_{i=1,i\neq j}^{K} (p_{i\bullet} - \alpha_i) = \alpha_j + \pi_j \sum_{i=1}^{K} (p_{i\bullet} - \alpha_i) = \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, \( p_{ij} = \delta_{ij}\tilde{\alpha}_j + (p_{\bullet j} - \tilde{\alpha}_j)\tilde{\pi}_j \), with \( \tilde{\alpha}_j = p_{\bullet j}\tilde{\Delta}_j, \quad p_{\bullet j} = \tilde{\alpha}_j + (1 - \tilde{\Delta})\tilde{\pi}_j \) and \( \tilde{\Delta} = \sum_{j=1}^{K} \tilde{\alpha}_j \). Since \( p_{ij} \) must take the same value in both models, \( p_{ij} = (p_{i\bullet} - \alpha_i)\pi_j = (p_{i\bullet} - \alpha_j)\pi_j \) for \( i \neq j \), that is, \( (p_{i\bullet} - \alpha_i)/\pi_i = (p_{i\bullet} - \alpha_j)/\pi_j \quad (\forall i \neq j) \). Assuming that \( K > 2 \), since the case of \( K = 2 \) requires special
treatment, the previous equality takes the constant value $\gamma$; and so

\[ p_i - \alpha_i = \gamma \tilde{\pi}_i \text{ and } p_i - \tilde{\alpha}_i = \gamma \pi_j \]  

(A3)

By adding the values in $i$ or $j$, we obtain

\[ 1 - \sum_{i=1}^K \alpha_i = 1 - \sum_{i=1}^K \tilde{\alpha}_i = \gamma, \text{ and } \Delta = \tilde{\Delta}. \]

In the case of $i = j$, we obtain

\[ \rho_{ij} = \alpha_i + (p_i - \alpha_i)\pi_i = \tilde{\alpha}_i + (p_i - \tilde{\alpha}_i)\tilde{\pi}_i, \text{ so } \alpha_i + \gamma \pi_i \tilde{\pi}_i = \tilde{\alpha}_i + \gamma \pi_i \pi_i \text{ and } \alpha_i = \tilde{\alpha}_i. \]

Substituting expression (A3) into (A2) and considering that $\gamma = 1 - \Delta$, we obtain

\[ \rho_{ij} = \delta_{ij}\alpha_i + (1 - \Delta)\pi_i \pi_j, \]  

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’:

\[ p_i = p_i \Delta_i + (1 - \Delta)\pi_i, \quad \tilde{\pi}_i = \frac{p_i \Delta_i}{p_i \Delta_i + (1 - \Delta)\pi_i} \text{ and } \hat{\pi}_i = \frac{p_i (1 - \Delta_i)}{1 - \Delta} \]  

(A4)

The first equality stems from the final statement in the first paragraph; the second equality from the fact that $\alpha_i = p_i \Delta_i = \tilde{\alpha}_i = p_i \tilde{\Delta}_i$; and the third equality from the fact that $\hat{\pi}_i = (p_i - \tilde{\alpha}_i)/(\gamma = (p_i - \alpha_i)/(\gamma = p_i(1 - \Delta_i))/(1 - \Delta)$. 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.

\[ a_i = p_i \Delta_i, \Delta = \sum_{i=1}^K p_i \Delta_i, \pi_{i2} = \pi_i, \pi_{i1} = p_i (1 - \Delta_i), p_i = \alpha_i + (1 - \Delta)\pi_i \]  

(A5)

Appendix 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

\[ \rho_{ijh} = \delta_{ijh}\alpha_i + (1 - \Delta)\pi_{i1}\pi_{j2}\pi_{h3}, \]

with $i, j, h = 1, 2, \ldots, K$ and $\Delta = \sum_{i=1}^K \alpha_i$. If $\tilde{\rho}_{ijh} = x_{ijh}/n$ are the observed proportions, then, except for one constant the logarithm of the likelihood is

\[ L = \sum_{i=1}^K \sum_{j=1}^K \sum_{h=1}^K \tilde{\rho}_{ijh} \log \rho_{ijh} = \sum_{i=1}^K \sum_{j=1}^K \sum_{h=1}^K \tilde{\rho}_{ijh}(1 - \delta_{ijh}) \log \rho_{ijh} = \sum_{i=1}^K \tilde{\rho}_{pii} \log \pi_i + \sum_{i=1}^K \tilde{\rho}_{pij} \log (1 - \Delta) + \sum_{i=1}^K \tilde{\rho}_{pij} \log \pi_j + \sum_{i=1}^K \tilde{\rho}_{pij} \log \pi_h. \]

By working it out we obtain

\[ L = \sum_{i=1}^K \tilde{\rho}_{pii} \log \pi_i + (1 - \tilde{\rho}) \log (1 - \Delta) + \sum_{i=1}^K \sum_{j=1}^K \sum_{h=1}^K \tilde{\rho}_{ijh} \log \pi_i \pi_j \pi_h, \]

where $\tilde{\rho} = \sum_{i=1}^K \tilde{\rho}_{pii}$ is the total proportion of agreements, $r = 1, 2, 3$ and $d_{ir}$ is the proportion of disagreements in category $i$ of rater $r$, that is,

\[ d_{ir} = \tilde{p}_{ir} - \tilde{\rho}_{iri}, d_{i2} = \tilde{p}_{iri} - \tilde{\rho}_{iri} \text{ and } d_{i3} = \tilde{p}_{iri} - \tilde{\rho}_{iri}. \]

Bearing in mind that $\tilde{p}_{iri} = \alpha_i + (1 - \Delta)\pi_i \pi_j \pi_h$, the following derivatives are obtained

\[ \frac{\partial \tilde{p}_{iri}}{\partial \alpha_i} = \delta_{iir} - \frac{\tilde{p}_{iri} - \alpha_i}{1 - \Delta}, \quad \frac{\partial \tilde{p}_{iri}}{\partial \pi_j} = \delta_{iir} \frac{\tilde{p}_{iri} - \alpha_i}{\pi_j}, \]

and the derivatives of $L$ are

\[ \frac{dL}{d\alpha_i} = \frac{\tilde{p}_{iri}}{p_i} - \frac{A}{1 - \Delta} \quad \text{with } A = 1 - \sum_{i=1}^K \tilde{p}_{iri} \alpha_i, \quad \frac{dL}{d\pi_j} = \frac{1}{\pi_j} \left\{ \tilde{d}_{ir} + \tilde{p}_{iri} - \tilde{p}_{iri} \alpha_i \right\} = A_{ir} \]

(B1)

Since the maximum likelihood estimates of $\alpha_i$ must verify that $dL/d\alpha_i = 0$, then $\tilde{p}_{iri}/p_{iri} = A(1 - \Delta)$ through the first expression of (B1). By substituting in the definition of $A, A = 1 - \Delta, \tilde{p}_{iri}/p_{iri} = 1$ and $\tilde{p}_{iri} = p_{iri}$. 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 $p_{iri}$ would be $\hat{p}_{iri} = \hat{\alpha}_i + (1 - \Delta)\hat{\pi}_{i1}\hat{\pi}_{i2}\hat{\pi}_{i3} = \hat{p}_{iri}$. Thus

\[ \hat{\alpha}_i = \tilde{p}_{iri} - (1 - \Delta)\hat{\pi}_{i1}\hat{\pi}_{i2}\hat{\pi}_{i3} \text{ and } \hat{p}_{iri} = \tilde{p}_{iri}, \]

(B2)
where \( \hat{\Delta} = \sum_{i=1}^{K} \hat{\alpha}_i \) is the maximum likelihood estimator of \( \Delta \). Note that \( \hat{\alpha}_i = \tilde{p}_{iii} \) if \( \exists r | \tilde{\pi}_{ir} = 0 \).

Substituting \( \tilde{p}_{iii} = p_{iii} \) in the second expression of (B1), \( \partial L / \partial \pi_{ir} = (\tilde{d}_{ir} + \tilde{p}_{iii} - \alpha_i) / \pi_{ir} = A_{ir} \). If the maximum likelihood estimators of \( \pi_{ir} \) have to verify that \( dL/dx_{ir} = 0 \) 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 \( \sum_{r=1}^{R} A_r \pi_{ir} = A_r = 1 - \Delta \). Therefore, \( A_r = A = 1 - \Delta \). Given that \( \tilde{p}_{iii} - \alpha_i = p_{iii} - \alpha_i = (1 - \Delta) \pi_{11} \pi_{12}, \) from the last expression (B1), we obtain

\[
\frac{\partial L}{\partial \pi_{ir}} = \frac{1}{\pi_{ir}} (\tilde{d}_{ir} + (1 - \Delta) \pi_{11} \pi_{12} \pi_{i3}) = 1 - \Delta \text{ or } (1 - \Delta) \pi_{ir} = \tilde{d}_{ir} + (1 - \Delta) \pi_{11} \pi_{12} \pi_{i3}. \tag{B3}
\]

These expressions yield the maximum likelihood estimators \( \hat{\Delta} \) and \( \hat{\pi}_{ir} \). Once the estimators are obtained, \( \tilde{d}_{i3} = \tilde{p}_{iii} - \tilde{p}_{ir} = (1 - \hat{\Delta}) \hat{\pi}_{i1} - (1 - \hat{\Delta}) \hat{\pi}_{i2} \hat{\pi}_{i3} \); thus, \( \tilde{d}_{i3} = \tilde{d}_{i1} \) from the last equality of (B3) and \( \tilde{p}_{ir} \). In general, \( \hat{\pi}_{ir} = \tilde{d}_{ir} \) and \( \hat{\pi}_{ir} = \tilde{d}_{ir} \).

The estimators of \( \Delta \) and \( \pi_{ir} \) can be obtained based on the expressions (B3). If \( \tilde{d}_{i3} = 0 \), then \( \partial L / \partial \pi_{i3} = (1 - \Delta) \pi_{11} \pi_{12} \leq 1 - \Delta, \) \( dL/d\pi_{i3} \leq 0 \), and \( \hat{\pi}_{i3} = 0 \); generally if \( \tilde{d}_{ir} = 0 \), then \( \hat{\pi}_{ir} = 0 \). Let \( \lambda_i = (1 - \Delta) \pi_{11} \pi_{12}, \) the last equality in expression (B3) indicates that \( \pi_{ir} = (\tilde{d}_{ir} + \lambda_i) / (1 - \Delta), \) and therefore, \( \lambda_i = (\tilde{d}_{i1} + \lambda_i) (\tilde{d}_{i2} + \lambda_i) (\tilde{d}_{i3} + \lambda_i) / (1 - \Delta)^2 \) (\( \forall i \)). If \( B = 1 - \Delta \), the following expressions must be solved for \( \lambda_i \):

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

where \( \lambda_i = 0 \) if \( \forall r | \tilde{d}_{ir} = 0 \). The condition \( g(B) = 0 \) is due to the fact that since \( \pi_{ir} = (\tilde{d}_{ir} + \lambda_i) / B \) and \( \sum_{i=1}^{K} \pi_{ir} = 1 \), then \( B \sum_{i=1}^{K} \pi_{ir} = D + \sum_{i=1}^{K} \lambda_i = B \). 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 \( \pi_{ir} = (\tilde{d}_{ir} + \lambda_i) / B \), 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):

\[
\hat{\pi}_{ir} = \frac{\lambda_i + \tilde{d}_{ir}}{B}, \hat{\alpha}_i = \tilde{p}_{ir} - \lambda_i, \hat{\Delta} = 1 - B \text{ and } \hat{\delta}_i = \frac{3\hat{\alpha}_i}{\tilde{p}_{ir} + \tilde{p}_{ir} + \tilde{p}_{ir}} = \frac{3\hat{\alpha}_i}{3\tilde{p}_i + \tilde{D}_i}, \tag{B5}
\]

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

\[
B = \frac{\tilde{d}_{ir}}{\hat{\pi}_{ir} - \prod_{r'=1}^{R} \hat{\pi}_{ir}}, \tag{B6}
\]

so that the values of \( \hat{\pi}_{ir} \) depend solely on the values of \( \hat{d}_{ir} \) and not on the values of \( \hat{\pi}_{ir} \).

By reason of expression (9),

\[
\hat{\Delta} = \frac{\hat{I}_e - \hat{I}_r}{1 - \hat{I}_r}, \text{ where } \hat{I}_r = \sum_{r=1}^{R} \prod_{r'=1}^{R} \hat{\pi}_{ir},
\]

and through the expression (5),

\[
\hat{\Delta} - \hat{\kappa}_{HR} = \frac{(1 - \hat{\Delta})(1 - \hat{\kappa}_{HR})}{(1 - \hat{I}_e)} (\hat{I}_e - \hat{I}_r), \tag{B7}
\]

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

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

The result of this is that \( \hat{I}_e = \sum_{i=1}^{K} \hat{\pi}_i R \) and \( \hat{I}_p = \sum_{i=1}^{K} \hat{\pi}_i R \). In these cases, the minimum (or maximum) value of \( \hat{I}_e \) is \( 1/K^{R-1} \) (or 1), which is reached when \( \hat{t}_i = 1/K \) (or when all the \( \hat{t}_i \) have a value of 0, except one which has a value of 1); it is similar with \( \hat{\pi}_i \), but now based on the probabilities \( \hat{\pi}_i \). When the marginal distributions of the proportions of responses are totally balanced, \( \hat{t}_i = 1/K (\forall i) \), \( \hat{I}_e \) 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, \( \hat{d}_i = \hat{D}/K (\forall i) \), \( B = \hat{D}/K (\hat{\pi}_i - \hat{\pi}_i R) (\forall i) \), \( \hat{\pi}_i = 1/K (\forall i) \), \( \hat{I}_\pi \) 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 \( \hat{I}_e \) approaches its maximum value. If \( \hat{t}_1 \) is very large, the rest of the values of \( \hat{t}_i \) will be small and \( \hat{I}_e \) will be close to 1; but that large marginal unbalance does not imply that the values \( \hat{d}_i \) 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 \( \hat{I}_e \) and, finally, \( \hat{\Delta} \) will be appreciably larger than \( \kappa_{HR} \).

**Appendix C. Variances of estimates when \( R > 2 \) or \( K > 2 \)**

Agresti [35] specifies the multivariate delta method in the case of a multinomial. If a function \( f = f(p_{ijh}) \) is estimated by \( \hat{f} = f(\hat{p}_{ijh}) \), with \( p_{ijh} = \{p_{ijh}\}, \hat{p}_{ijh} = \{\hat{p}_{ijh}\} \) 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

\[
V(\hat{f}) = \left[ \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{h=1}^{K} f_{ijh}^2 \hat{p}_{ijh} - \left( \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{h=1}^{K} f_{ijh} \right)^2 \right] / n. \tag{C1}
\]

A similar notation to the one used for the observed proportions (\( \hat{p}_{ijh}, \hat{d}_{ir}, \hat{p}_i, \hat{t}_{ir} \)) 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

\[
\frac{\lambda_i}{\lambda_i + d_{ir}} = \frac{P_i}{\pi_{ir}} \text{ and } \sum_{r=1}^{R} \frac{\lambda_i}{\lambda_i + d_{ir}} = P_i I_i \text{ with } P_i = \sum_{r=1}^{R} \pi_{ir} \text{ and } I_i = \sum_{r=1}^{R} \pi_{ir}^{-1}. \tag{C2}
\]

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

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

\[
\Delta'_{ijh} = \sum_{t=1}^{K} \lambda'_{t(ijh)} + D'_{ijh} \sum_{t=1}^{K} \lambda'_{t(B)} - 1. \tag{C3}
\]

For the first equality of (11), \( 2\log B = \sum_{t=1}^{R} \log(\lambda_t + d_{ir}) - \log \lambda_i \); deriving this expression, we obtain \( \lambda'_{t(B)} = 2X_i \), and the denominator of expression (C3) will be \( 2X - 1 \). 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
Finally, if \( \beta_i \) then \( \sum_{r=1}^{R} (P_r / \pi_{tr}) (\partial d_r / \partial p_{ijh}) \) and \( \lambda'_{t(ih)}[1-P_r I_t] \) is 
\[ \lambda'_{t(ih)}[1-P_r I_t] = \sum_{r=1}^{R} (P_r / \pi_{tr}) (\partial d_r / \partial p_{ijh}) \text{ and } \lambda'_{t(ih)} = -(1-\delta_{ijh})X_t(\pi_{t11}) + (\delta_{ijh} / \pi_{t12}) \text{; therefore, } \]
\[ \sum_{i=1}^{K} \lambda'_{t(ih)} = \lambda'_{t(ih)} + \lambda'_{t(ih)} + \lambda'_{h(ih)} = -(1-\delta_{ijh})(1-(X_t / \pi_{t11}) + (X_t / \pi_{t12}) + (X_t / \pi_{t13})). \]
Replacing everything in expression (C3),
\[ \Delta'_{ijh} = f_{ijh} = \frac{1 - \delta_{ijh}}{2X - 1} \left[ 1 - \left( \frac{X_j}{\pi_{t1}} + \frac{X_h}{\pi_{t3}} \right) \right], \]
where the number 2 will be \((R-1)\). To apply expression (C1), we must consider that
\[ (1-\delta_{ijh})p_{ijh} = (1-\delta_{ijh})(\delta_{ijh} \omega_t + B \pi_{t12} \pi_{t13}) = (1-\delta_{ijh})B \pi_{t12} \pi_{t13} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{h=1}^{K} f_{ijh}P_{ijh} = -B, \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{h=1}^{K} f_{ijh}P_{ijh} = B(3X-1)/(2X-1) \text{-which in general will be } B(RX-1)/[(R-1)X-1]. \]
Therefore, its variance is similar to the variance of expression (15) and (16).

\[ V(\Delta^*) = \frac{1}{n} \left[ \alpha_t(1-\alpha_t) + (1-\Delta)X_t \left( \frac{R-1}{(R-1)X-1} - 1 \right) \right]. \]

Finally, if \( \hat{f} = \hat{S}_t \) then, by expression (B5), \( f = S_t = 3\alpha_t / N_t \) and \( f_{ijh} = \partial S_t / \partial p_{ijh} = [3(N_t) \times \Delta'_{t(ij)} - \alpha_t(\delta_{ij} + \delta_{ij} + \delta_{ij}) / N_t] \) with \( N_t = p_{\bullet t} + p_{\bullet \bullet t} + p_{\bullet \bullet t} = 3p_t + D_t \). We obtain \( \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{h=1}^{K} f_{ijh}P_{ijh} = 0 \) and \( \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{h=1}^{K} f_{ijh}^2P_{ijh} = nV(S_t) \) by expression (C1), with
\[ V(\hat{S}_t) = \frac{1}{n} \left( \frac{3}{N_t} \right)^2 \left[ \frac{2X_t}{2X-1} - 1 \right] BX_t + \alpha_t - \frac{(2 \times 3 - 1)\alpha_t^2}{N_t} + 2\left( \frac{\alpha_t}{N_t} \right)^2 (p_{\bullet \bullet t} + p_{\bullet \bullet t} + p_{\bullet \bullet t}) \].

Generalizing to any value of \( R, N_t = R p_t + D_t = R \alpha_t + (1-\Delta) \sum_{r=1}^{R} \pi_{tr} \), and
\[ V(\hat{S}_t) = \frac{R^2}{nN_t^2} \left[ BX_t \left( \frac{R-1}{(R-1)X-1} - 1 \right) + \alpha_t(1-\hat{S}_t) \left( \frac{R-1}{R} - \hat{S}_t \right) \right] + B \left( \frac{S_t}{R} \right)^2 \left( \left( \sum_{r=1}^{R} \pi_{tr} \right)^2 - \left( \sum_{r=1}^{R} \pi_{tr}^2 \right) \right) \].

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.

**Appendix D. Variances in the special case of** \( R = K = 2 \)

As indicated in Section 4.2, the observed proportions verify the equalities \( \bar{p}_{13} = \bar{p}_{33} = \bar{p}_{33} = \bar{d}_{11} = \bar{d}_{22}, \bar{d}_{12} = \bar{d}_{21}, \) and \( \bar{d}_{13} = \bar{d}_{23} = 2\bar{p}_{33}. \) If we take the first expression (10), we obtain \( \pi_{11} = \pi_{22}, \) \( \pi_{12} = \pi_{21}, \) \( \pi_{13} = \pi_{23} \), so \( \hat{X}_1 = \hat{X}_2 \) and \( \hat{X}_1 = \pi_{23}^3 / (2\hat{\pi}_{33} - 1). \) Determining \( V(\hat{S}_t) \) does not pose any problem, as the consistency \( \hat{S}_t \) 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 \( f = \alpha_t^*/(1-p_{3*}) \) and \( f_{ij} = (\partial f/\partial \alpha_t) \times (\partial \alpha_t/\partial p_{ij}) + (\partial f/\partial p_{3*}) \times (\partial p_{3*}/\partial p_{ij}) = [\partial \alpha_t/\partial p_{ij} + \alpha_t^* \delta_{ij}]/(1-p_{3*}) \), 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 \( \sum_{i=1}^{K} \sum_{j=1}^{K} f_{ij} p_{ij} = \alpha_t^*/(1-p_{3*}) \), \( \sum_{i=1}^{K} \sum_{j=1}^{K} f_{ij}^2 p_{ij} = [(1-\Delta)X_i(X_i/(X-1) - 1) + (1-p_{3*})\alpha_t^*(1-\alpha_t^*) + (\alpha_t^*)^2]/(1-p_{3*})^2 \) and, through expression (C1), we obtain the following variance that yields the second expression (16)

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

If \( \hat{f} = \hat{\Delta}^* \), then \( f = \Delta/(1-p_{3*}) \) and \( f_{ij} = (\partial f/\partial \Delta) \times (\partial \Delta/\partial p_{ij}) + (\partial f/\partial p_{3*}) \times (\partial p_{3*}/\partial p_{ij}) = [\Delta_{ij} - \Delta^{3*}_{ij} + \Delta^*\delta_{ij}]/(1-p_{3*}) \), 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 \( \sum_{i=1}^{K} \sum_{j=1}^{K} f_{ij} p_{ij} = -(1-\Delta^*)/(1-p_{3*}) \), \( \sum_{i=1}^{K} \sum_{j=1}^{K} f_{ij}^2 p_{ij} = [(R(1-X_3)(X-X_3)/(X-1)) + (1-\Delta^*)^2]/(1-p_{3*})^2 \) and, through expression (C1), we obtain the following variance that yields expression (17)

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