PROPOSAL

NOMINAL AGREEMENT MEASURES AMONG MANY RATERS

The following is based on the paper Martín Andrés and Álvarez Hernández (2022)

1. Measures of agreement based on *Multi-rater Delta* model

If $R \geq 2$ raters independently classify n subjects in K categories, a data matrix $\{y_{sr}\}$ is obtained, with $s=1, 2, \ldots, n$, $r=1, 2, \ldots, R$ and $y_{sr}=1, 2, \ldots, K$, in which $y_{sr}=i$ when the rater r classifies subject s into category i. The most common 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 (see Table 1a). With this classification, $\left\{x_{i_1 i_2 \ldots i_R}\right\}$ is a multinomial random variable of sample size n and probabilities $\left\{p_{i_1 i_2 \ldots i_R}\right\}$. The "crude" agreement (without random correction) for these data is $\Sigma_i \, \overline{p}_{i \ldots i} = 0.610$.

The *Multi-rater Delta* model assumes that these probabilities are given by

$$p_{i_1 i_2 \ldots i_R} = \delta_{i_1 i_2 \ldots i_R} \alpha_{i_1} + (1-\Delta) \prod_{r=1}^{r=R} \pi_{i_r r} \quad \text{with } \Delta = \Sigma \alpha_i \,, \qquad \textbf{(1)}$$

where $i_r=1, 2, \ldots, K$, $\Delta \leq 1$ and $\alpha_i \leq 1$ and $0 \leq \pi_{i_r r} \leq 1$ are the parameters of the model. In the *Multi-rater Delta* model the following response model is assumed. When the R raters face a given subject, they all recognize it as category i with intensity α_i; when they recognize it, they classify it as type i; when they do not recognize it, they do so with intensity $1-\Delta$, classify it randomly and independently with the probability distributions $\{\pi_{ir}\}$, $r=1, 2, \ldots, R$, respectively. The parameters of interest are:

α_i: the proportion of agreements in category i that are not due to chance,

Δ: the total proportion of agreements that are not due to chance = the overall degree of agreement,

$\mathcal{S}_i = \dfrac{R\alpha_i}{p_{i_1 \bullet \ldots \bullet} + \ldots + p_{\bullet \ldots \bullet i_R}} = \dfrac{R\alpha_i}{R\alpha_i + (1-\Delta)\sum_{r=1}^{R} \pi_{ir}}$: the degree of agreement in category i,

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.

To make inferences with the *Multi-rater Delta* model, only some of the observed relative frequencies $\overline{p}_{i_1 i_2 \ldots i_R} = x_{i_1 i_2 \ldots i_R} / n$ are needed: the observed proportions of agreements in category i, $\overline{p}_i = \overline{p}_{ii \ldots i} = \#\{s \mid y_{s1} = \ldots = y_{sR} = i\}/n$, and the observed proportions of disagreements by rater r in category i, $\overline{d}_{ir} = \overline{t}_{ir} - \overline{p}_i$, where $\overline{t}_{ir} = \sum_{i_1=1}^{K} \ldots \sum_{i_{r-1}=1}^{K} \sum_{i_{r+1}=1}^{K} \ldots \sum_{i_R=1}^{K} \overline{p}_{i_1 \ldots i_r = i \ldots i_R} = \#\{s \mid y_{sr}=i\}/n$ is the total proportion of responses i of rater r. Based on these proportions, the total proportion of agreements

$\bar{p} = \sum_{i=1}^{K} \bar{p}_i$, the total proportion of disagreements $\bar{D} = \sum_{i=1}^{K} \bar{d}_{ir}$ (which is the same for all raters), and the total proportion of disagreements in category i $\bar{D}_i = \sum_{r=1}^{R} \bar{d}_{ir}$ are determined (see Table 1b). Based on the definitions, $\bar{t}_{ir} = \bar{p}_i + \bar{d}_{ir}$, $1 = \bar{p} + \bar{D}$, $\bar{D} = \sum_{i=1}^{K} \bar{D}_i / R$ and $\bar{D}_i \le (R-1)\bar{D}$.

Once these proportions are known, the parameter estimates of interest are as follows:

$$\hat{\pi}_{ir} = \frac{\lambda_i + \bar{d}_{ir}}{B},\ \hat{\alpha}_i = \bar{p}_i - \lambda_i,\ \hat{\Delta} = 1 - B \text{ and } \hat{\mathcal{S}}_i = \frac{R\hat{\alpha}_i}{\bar{N}_i} = \frac{R\hat{\alpha}_i}{R\hat{\alpha}_i + (1-\hat{\Delta})\sum_{r=1}^{R}\hat{\pi}_{ir}}, \quad \textbf{(2)}$$

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

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

with the exception that $\lambda_i = 0$ when $\bar{d}_{ir} = 0$ for some rater r. In the particular case of sample independence, 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 (3); thus, $\hat{\alpha}_i = \hat{\Delta} = 0$.

After obtaining the parameter estimation and measures of agreement of the *Multi-rater Delta* model (Δ, α_i, π_{ir} and $\mathcal{S}_i$), their variances must be obtained to make inferences about the measures of agreement. The estimated variances are:

$$\hat{V}(\hat{\Delta}) = \frac{1-\hat{\Delta}}{n}\left\{\hat{\Delta} + \frac{\hat{X}}{(R-1)\hat{X} - 1}\right\}, \quad \textbf{(4)}$$

$$\hat{V}(\hat{\alpha}_i) = \frac{1}{n}\left[\hat{\alpha}_i(1-\hat{\alpha}_i) + (1-\hat{\Delta})\hat{X}_i\left\{\frac{(R-1)\hat{X}_i}{(R-1)\hat{X} - 1} - 1\right\}\right] \text{ and} \quad \textbf{(5)}$$

$$\hat{V}(\hat{\mathcal{S}}_i) = \frac{R^2}{n\bar{N}_i^2}\left[\begin{array}{l} n\hat{V}(\hat{\alpha}_i) - \hat{\alpha}_i(1-\hat{\alpha}_i) + \hat{\alpha}_i(1-\hat{\mathcal{S}}_i)\left\{1 - \frac{R-1}{R}\hat{\mathcal{S}}_i\right\} + \\ + \frac{(1-\hat{\Delta})\hat{\mathcal{S}}_i^2}{R^2}\left\{\left(\sum_{r=1}^{R}\hat{\pi}_{ir}\right)^2 - \left(\sum_{r=1}^{R}\hat{\pi}_{ir}^2\right)\right\} \end{array}\right], \quad \textbf{(6)}$$

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 $\hat{X} = \sum_{i=1}^{K}\hat{X}_i$ (see Table 1c). 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 $B = \infty$ because then there exists some $\hat{\pi}_{ir} = 0$. In these cases, variances can be estimated if the calculations are performed for the data increased by 0.5; thus, the new sample size is $n + K^R/2$, and the new frequencies observed are $\bar{p}_{i_1 i_2 \ldots i_R} = (x_{i_1 i_2 \ldots i_R} + 0.5)/(n + K^R/2)$.

When only two raters and two categories exist, the *Multi-rater Delta* model has more unknown parameters (α_1, α_2, π_{11} and π_{12}) than free cells to take values (three). In this case, the following solution by Martín Andrés and Femia Marzo (2004, 2005) can be adopted. 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 0.5, estimate the parameters as performed in the previous section, and redefine the measures of agreement without considering the third dummy category. Let $\overline{p}_i$ and $\overline{d}_{ir}$ be the new observed frequencies and α_i, π_{ir} and α be the parameters of the *Alpha* 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 are

$$\hat{\alpha}_i^* = \frac{\hat{\alpha}_i}{1-\overline{p}_{3\bullet}},\ \hat{\mathrm{V}}\left(\hat{\alpha}_i^*\right) = \frac{1}{n\left(1-\overline{p}_{3\bullet}\right)^2}\left[\left(1-\hat{\Delta}\right)\hat{X}_i\left\{\frac{\hat{X}_i}{\hat{X}-1}-1\right\}+\left(1-\overline{p}_{3\bullet}\right)\hat{\alpha}_i^*\left(1-\hat{\alpha}_i^*\right)\right], \qquad \textbf{(7)}$$

$$\hat{\Delta}^* = \hat{\alpha}_1^* + \hat{\alpha}_2^*,\ \hat{\mathrm{V}}\left(\hat{\Delta}^*\right) = \frac{1}{n\left(1-\overline{p}_{3\bullet}\right)^2}\left[\left(1-\hat{\Delta}\right)\left(1-\hat{X}_3\right)\frac{\hat{X}-\hat{X}_3}{\hat{X}-1}+\left(1-\overline{p}_{3\bullet}\right)\hat{\Delta}^*\left(1-\hat{\Delta}^*\right)\right], \qquad \textbf{(8)}$$

$$\hat{\mathcal{S}}_i^* = \frac{2\hat{\alpha}_i}{\overline{N}_i},\ \hat{\mathrm{V}}\left(\hat{\mathcal{S}}_i^*\right) = \frac{4}{n\overline{N}_i^2}\left[\left(1-\hat{\Delta}\right)\hat{X}_i\left\{\frac{\hat{X}_i}{\hat{X}-1}-1\right\}+\hat{\alpha}_i\left\{1-\frac{3\hat{\alpha}_i}{\overline{N}_i}+\frac{2\hat{\alpha}_i^2}{\overline{N}_i^2}+\frac{2\hat{\pi}_{i1}\hat{\pi}_{i2}\left(1-\hat{\Delta}\right)\hat{\alpha}_i}{\overline{N}_i^2}\right\}\right], \qquad \textbf{(9)}$$

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

2. Measures of agreement based on *kappa*

A common measure of agreement when only two raters (R=2) exist is the Cohen's *kappa* coefficient (κ_C) (1960) estimated by $\hat{\kappa}_C = \left(\overline{I}_0 - \overline{I}_e\right)/\left(1-\overline{I}_e\right)$, where $\overline{I}_o = \sum_{i=1}^{K}\overline{p}_{ii}$ is the observed agreement index, and $\overline{I}_e = \sum_{i=1}^{K}\overline{p}_{i\bullet}\overline{p}_{\bullet i}$ is the expected agreement index under the assumption of independence between the classifications of the two raters. The κ_C coefficient can be generalized to the case of multiple raters (R≥2) in several ways, depending on how the phrase "an agreement occurs" is interpreted.

Hubert (1977) makes the following interpretation, which is similar to Cohen's definition: "an agreement occurs if and only if all raters agree on the categorization of an object" or DeMoivre's definition of agreement (R-wise). In this case, Hubert's *kappa* (κ_{HR}) is estimated by $\hat{\kappa}_{HR} = \left(\overline{I}_o - \overline{I}_e\right)/\left(1-\overline{I}_e\right)$, with $\overline{I}_o = \sum_{i=1}^{K}\overline{p}_i$ and $\overline{I}_e = \sum_{i=1}^{K}\Pi_{r=1}^{R}\overline{t}_{ir}$, such that $\hat{\kappa}_{HR}=\hat{\kappa}_C$ when R=2. The estimated variance of $\hat{\kappa}_{HR}$ is (Martín Andrés and Álvarez Hernández 2020)

$$\hat{V}\left(\hat{\kappa}_{HR}\right)=\frac{\hat{U}+\hat{V}-\hat{W}}{n\left(1-\overline{I}_e\right)^2} \quad \text{with} \quad \begin{cases} \hat{U}=\sum_{i=1}^{K}\overline{p}_i\left[1-\left(1-\hat{\kappa}_{HR}\right)\sum_{r=1}^{R}\overline{T}_{ir}\right]^2 \\ \hat{V}=\left(1-\hat{\kappa}_{HR}\right)^2\sum_{i_1=1}^{K}\sum_{i_2=1}^{K}\cdots\sum_{i_R=1}^{K}\left(1-\delta_{i_1 i_2 \ldots i_R}\right)\overline{p}_{i_1 i_2 \ldots i_R}\left(\sum_{r=1}^{R}\overline{T}_{i_r r}\right)^2 \\ \hat{W}=\left[\left(R-1\right)\left(1-\hat{\kappa}_{HR}\right)\overline{I}_e-\hat{\kappa}_{HR}\right]^2 \end{cases} \quad \textbf{(10)}$$

where $\overline{T}_{ir}=\prod_{r=1}^{R}\prod_{r'=1,\,r'\neq r}^{R}\overline{t}_{ir'}$ (see Tables 1d and f). If sample independence exist, that is, if $\overline{p}_i=\prod_{r=1}^{R}\overline{t}_{ir}$, then $\overline{I}_o=\overline{I}_e$ and $\hat{\kappa}_H$=1. It can be seen that $\hat{V}\left(\hat{\kappa}_{HR}\right)=\hat{V}\left(\hat{\kappa}_C\right)$ when *R*=2, where $\hat{V}\left(\hat{\kappa}_C\right)$ is the value of Fleiss *et al.* (1969).

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 Hubert (1977) or pairwise definition of agreement. However, *kappa* coefficients can vary from one author to another, depending on the definition of I_e. The most traditional definition is the definition by Fleiss (1971), which yields the Fleiss' *kappa* coefficient (κ_F), estimated by

$$\hat{\kappa}_F=1-\frac{nR^2-\sum_{s=1}^{n}\sum_{i=1}^{K}R_{si}^2}{nR\left(R-1\right)\left\{1-\sum_{i=1}^{K}R_i^2\right\}}, \quad \textbf{(11)}$$

where $R_{si}=\#\{r\,|\,y_{sr}=i\}$=0, 1, …, *R* is the number of raters that classify subject *s* in category *i* and $R_i=R_{\bullet i}/nR=\overline{p}_i+\overline{D}_i/R$ is the total proportion of responses *i* (any rater) (see Tables 1e and f). The estimated variance of $\hat{\kappa}_F$ is (Schouten 1982, according to Vanbelle 2019)

$$\hat{V}\left(\hat{\kappa}_F\right)=\frac{\sum_s\left[\left(1-I_e\right)I_{o,s}-2\left(1-I_o\right)I_{e,s}-C\right]^2}{n^2\left(1-I_e\right)^4}, \quad \textbf{(12)}$$

where

$$\begin{cases} I_o=\dfrac{\sum_s I_{o,s}}{n} \\ I_e=\dfrac{\sum_s I_{e,s}}{n} \\ C=I_oI_e-2I_e+I_o \end{cases} \quad \text{and} \quad \begin{cases} I_{o,s}=\dfrac{\sum_i R_{si}^2-R}{R\left(R-1\right)} \\ I_{e,s}=\dfrac{\sum_i R_{si}R_{\bullet i}}{nR^2} \end{cases} \quad \textbf{(13)}$$

The Fleiss' *kappa* coefficient does not coincide with the Cohen's *kappa* coefficient when *R=2*. Hubert provided the following pairwise definition that matches Cohen's *kappa* coefficient when *R* = 2. Hubert's *kappa* (κ_{H2}) (2-wise) is estimated by:

$$\hat{\kappa}_{H2} = 1 - \frac{R^2 - \left(\sum_{s=1}^{n}\sum_{i=1}^{K} R_{si}^2 / n\right)}{R(R-1) - 2\sum_{i=1}^{K}\sum_{r=1}^{R}\sum_{r'=r+1}^{R} \bar{t}_{ir}\bar{t}_{ir'}}, \qquad \textbf{(14)}$$

REFERENCES

Cohen J (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*, 37-46. DOI: 10.1177/001316446002000104.

Dillon WR and Mulani N (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research 19*, 438-458. DOI: 10.1207/s15327906mbr1904_5.

Fleiss JL (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*, 378-382. DOI: 10.1037/h0031619.

Fleiss JL, Cohen J and Everitt BS (1969). Large Sample Standard Errors of Kappa and Weighted Kappa. *Psychological Bulletin 72*, 323-327. DOI: 10.1037/h0028106.

Hubert L (1977). Kappa revisited. *Psychological Bulletin 48(2)*, 289-297. DOI: 10.1037/0033-2909.84.2.289.

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*, 1-19. DOI: 10.1348/000711004849268.

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. DOI: 10.1191/0962280205sm412oa.

Martín Andrés, A. and Álvarez Hernández, M. (2022). Multi-rater delta: extension to many raters of the measure delta of nominal agreement. *Journal of Statistical Computation and Simulation 9 (2)*, 1877-1897. (published on line: 20 Dec 2021). DOI:10.1080/00949655.2021.2013485.

Martín Andrés, A. and Álvarez Hernández, M. (2020). Hubert's multi-rater kappa revisited. *British Journal of Mathematical and Statistical Psychology 73 (1)*, 1-22. DOI: 10.1111/bmsp.12167.

Schouten HJA (1982). Measuring pairwise agreement among many observers. ii. Some improvements and additions. *Biometrical Journal 24 (5)*, 431-435.

Vanbelle S (2019). Asymptotic variability of (multilevel) multirater kappa coefficients. *Statistical Methods in Medical Research 28 (10-11)*, 3012-3026.

Table 1

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*
Rater 1	*1*	56	1	0	5	3	0	0	0	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 1(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) Estimation of the parameters and measures of the degree of agreement 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*	*r*=3	*(degree of agreement in class i)*
1	0.3320	0.1564	0.5084	0.2647	0.7040±0.0460
2	0.0741	0.6343	0.2823	0.5937	0.2462±0.1011
3	0.1435	0.2093	0.2093	0.1416	0.6306±0.0668
Overall degree of agreement ($\hat{\Delta}\pm SE$)					0.5496±0.0462

(d) Data needed to estimate Hubert's *kappa* (DeMoivre or *R*-wise), which are obtained from Table 1(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

(e) Data needed to estimate Fleiss' *kappa*, which are obtained from Table 1(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)*
	ω=1	ω=2	ω=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= $\sum_{i=1}^{3} R_{\bullet i}=\sum_{\omega=1}^{3}\omega s_{\bullet\omega}$

(f) Summary of the different measures of degree of agreement for the data in Table 1(a)

Raw	0.610
Multi-rater delta	0.550
Hubert's kappa (*R-wise*)	0.547
Hubert's kappa (*2-wise*)	0.581
Fleiss' kappa	0.578