LIFE: International Journal of Health and Life-Sciences
ISSN 2454-5872

Xie et al.

Volume 3 Issue 3, pp.1- 15

Date of Publication: 16th November 2017

DOI-https://dx.doi.org/10.20319/lijhls.2017.32.115

This paper can be cited as: Xie, Z., Gadepalli, C. & Cheetham, B. M. G. (2017). Reformulation and Generalisation of the Cohen and Fleiss Kappas. LIFE: International Journal of Health and Life-Sciences, 3(3), 01-15.

REFORMULATION AND GENERALISATION OF THE COHEN AND FLEISS KAPPAS

Zheng Xie
School of Engineering, University of Central Lancashire, Preston, United Kingdom
zxie2@ uclan.ac.uk

Chaitanya Gadepalli
University Department of Otolaryngology, Central Manchester University Hospitals Foundation Trust and University of Manchester Academic Health Science Centre, Manchester, United Kingdom

Barry M.G. Cheetham
School of Computer Science, University of Manchester, Manchester, United Kingdom

Abstract

The assessment of consistency in the categorical or ordinal decisions made by observers or raters is an important problem especially in the medical field. The Fleiss Kappa, Cohen Kappa and Intra-class Correlation (ICC), as commonly used for this purpose, are compared and a generalised approach to these measurements is presented. Differences between the Fleiss Kappa and multi-rater versions of the Cohen Kappa are explained and it is shown how both may be applied to ordinal scoring with linear, quadratic or other weighting. The relationship between quadratically weighted Fleiss and Cohen Kappa and pair-wise ICC is clarified and generalised to multi-rater assessments. The AC_1 coefficient is considered as an alternative measure of consistency and the relevance of the Kappas and AC_1 to measuring content validity is explored

Keywords

Assessment of consistency and content validity, Fleiss Kappa, Cohen Kappa, ICC, Gwet's AC_1 coefficient, Multi-rater assessments, CVI

1. Introduction

A common problem in medicine and other fields is that of quantifying the consistency of decisions made by the same observers at different times or by different observers of the same phenomena (Gwet, 2014). The observers may be termed 'raters' and the decisions may be diagnoses of medical conditions or the severity of such conditions. The decisions may be 'categorical' such as 'improved', 'unchanged' or 'worse'. Or they may be 'ordinal' which means that they are 'scores', such as 0, 1, 2 and 3. Statistical techniques may be employed to assess intra-rater and inter-rater consistency. A well-known measure of correlation that may be considered for comparisons of ordinal (numerical) scoring is the Pearson correlation coefficient (Lee & Nicewander, 1998). Given N subjects scored $\{A(i)\}_{1,N}$ by rater A and $\{B(i)\}_{1,N}$ by rater B, it is defined as follows:

$$PC = \frac{\sum_{i=1}^{N}(A(i)-m_A)(B(i)-m_B)}{\sqrt{\sum_{i=1}^{N}(A(i)-m_A)^2\sum_{i=1}^{N}(B(i)-m_B)^2}} \quad (1)$$

where m_A and m_B are the arithmetic means of $\{A(i)\}_{1,N}$ and $\{B(i)\}_{1,N}$ respectively. As explained by Bland and Altman (Bland & Altman, 1986), Pearson Correlation is not normally appropriate for comparing pairs of raters as it takes into account only variations about the mean for each rater. A rater whose scores are consistently larger or smaller than those of another rater may appear perfectly correlated (PC=1) with that rater. An alternative is the 'intra-class correlation' coefficient (ICC) (Koch, 1982) which may be applied in its original form (Rödel, 1971) to measure intra-rater consistency, the inter-rater consistency of pairs of raters and also the inter-rater consistency of groups of raters. Given the set of scores referred to above, pair-wise ICC is defined as:

$$ICC = \frac{\sum_{i=1}^{N}(A(i)-m)(B(i)-m)}{0.5\left(\sum_{i=1}^{N}(A(i)-m)^2+\sum_{i=1}^{N}(B(i)-m)^2\right)} \quad (2)$$

where $m = (m_A + m_B)/2$. *ICC* compares differences between the scores of each rater and a 'pooled' arithmetic mean, *m*, computed over the scores of both raters. *ICC* is more indicative of rater consistency than Pearson Correlation. There are other versions of *ICC* (Müller & Büttner, 1994).

A simple measure of consistency is the 'proportion of agreement' (P_o). This just tells us how many times the decisions or scores of two raters agree. However, P_o reflects neither the magnitudes of any differences in ordinal scores nor the possibility of agreement by chance. With an even spread of decisions between four categories or scores, P_o would be 0.25 (or 25%) with purely random scoring, and with an uneven spread, P_o could be even greater by chance.

2. Cohen Kappa

Cohen Kappa is widely used for measuring the consistency of scores produced by a pair of raters *A* and *B*. In its original form (Cohen, 1960), it is applicable to categorical decisions, though it may be applied also to ordinal scoring if the scores are considered as labels. It is defined as follows:

$$K = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

where P_o is as defined above, and P_e is the estimated probability of agreement 'by chance' given the distribution of scores by raters *A* and *B*. When this form of Cohen Kappa is applied to ordinal scores, any difference is considered equally serious, regardless of its magnitude. It is useful to re-express the formula, in terms of disagreement as follows:

$$K = 1 - \frac{1 - P_o}{1 - P_e} = 1 - \frac{D_o}{D_e} \tag{4}$$

where $D_o = 1 - P_o$ is the proportion of actual scores that disagree. This may be considered a 'cost of actual disagreement'. $D_e = 1 - P_e$ is considered to be the estimated cost of 'by chance' disagreement given the actual distribution of scores. The 'weighted form of Cohen Kappa' (Cohen, 1968) allows the costs of actual disagreement and estimated 'by chance' disagreement between ordinal scores to be weighted according to the degree of the disagreement. To do this, D_o and D_e are redefined as follows:

$$D_o = \frac{1}{N}\sum_{i=1}^{N} C(A(i), B(i)) \tag{5}$$

$$D_e = \frac{1}{N^2}\sum_{i=1}^{K}\sum_{j=1}^{K} A_i B_j C(\alpha(i),\alpha(j)) \quad (6)$$

where A_i is the number of subjects which rater A scores as $\alpha(i)$, B_j is the number of subjects which rater B scores as $\alpha(j)$, and K is the number of possible scoring categories or numerical scores. The K scoring categories or numerical scores are denoted by $\alpha(1)$, $\alpha(2)$, ..., $\alpha(K)$. Cohen (Cohen, 1968) assumes that $\alpha(i) = i$ for $i = 1, 2, ..., K$. The cost-function C may be defined arbitrarily. If it is defined as follows:

$$C(a,b) = \begin{cases} 1 : a \neq b \\ 0 : a = b \end{cases} \quad (7)$$

the resulting 'weighted Cohen Kappa' becomes identical to the original unweighted Cohen Kappa (*UwCK*). For 'linearly weighted Cohen Kappa' (*LwCK*), the cost-function is:

$$C(a,b) = |a-b| \quad (8)$$

and for 'quadratically weighted Cohen Kappa'(*QwCK*) ,

$$C(a,b) = (a-b)^2 \quad (9)$$

There are many other possible cost-functions that may be considered, but these three are of special interest. The key feature of all forms of Cohen Kappa is that they aim to take into account the probability of agreement or disagreement 'by chance' given the distribution of decisions or scores produced by each rater. Equation (6) may be simplified and re-expressed as:

$$D_e = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i),B(j)) \quad (10)$$

Therefore we have a general formula for all forms of Cohen Kappa for pairs of raters:

$$Kappa = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} C(A(i),B(i))}{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i),B(j))} \quad (11)$$

A *Kappa* value of 1 indicates perfect consistency, and values in the ranges (0.8,1), (0.6,0.8), (0.4,0.6), (0.2,0.4), (0,0.2) are considered (Viera & Garrett, 2005) to indicate 'almost perfect', 'substantial', 'moderate', 'fair' and 'slight' consistency, respectively.

The way that the probability of agreement 'by chance' is estimated by Cohen Kappa has been strongly questioned by Gwet and others (Gwet, 2014). Alternatives to the Cohen Kappa have been proposed, such as the Aickin Alpha coefficient and the Gwet AC_1 coefficient (Gwet,

2014). There appear to be good reasons for adopting these newer coefficients, but at present the Cohen and Fleiss Kappas are more widely used.

The Kappa measures of consistency are related to the 'Content Validity Index' (CVI) measurements widely used in nursing and health care (Polit, & Beck 2006). Although the definitions of content validity are more specific than rater consistency, there are obvious similarities in that these two types of measurements are derived by panels of raters or experts. CVI measurements are used to validate questionnaires as used for health behaviour and need assessment (Kitreerawutiwo & Mekrungrongwong, 2015), and assessing the knowledge of caregivers (Sukron & Phutthikhamin, 2016). It has been suggested that a form of the Cohen Kappa can be an appropriate supplement, if not a substitute for CVI coefficients in view of its approach to the possibility of agreement by chance (Polit, & Beck 2006).

3. Relationship between ICC and Cohen Kappa

In 1973, Fleiss and Cohen (Fleiss & Cohen, 1973) established that there is 'equivalence' between quadratically weighted Cohen Kappa (*QwCK*) and *ICC* when "the systematic variability between raters is included as a component of variability". This equivalence may be expressed more directly as follows:

$$ICC = \frac{(1/N)\sum_{i=1}^{N} A(i)\times B(i) \; - \; m^2}{(0.5/N)\sum_{i=1}^{N}(A(i)^2 + B(i)^2) \; - \; m^2} \tag{12}$$

$$QwCK = \frac{(1/N)\sum_{i=1}^{N} A(i)\times B(i) \; - \; g^2}{(0.5/N)\sum_{i=1}^{N}(A(i)^2 + B(i)^2) \; - \; g^2} \tag{13}$$

where

$$m = \frac{m_A + m_B}{2} \quad \& \quad g = \sqrt{m_A \times m_B} \tag{14}$$

It follows that the difference between pair-wise *ICC* and *QwCK* depends only on the difference between the arithmetic m and the geometric mean g of m_A and m_B. These two means are usually close but not necessarily identical.

4. Fleiss Kappa

Both versions of the Cohen Kappa are applicable to pairs of raters. The Fleiss Kappa (Fleiss, 1971) is defined for measuring the agreement among more than two 'categorical' raters. A Fleiss Kappa of 1 indicates perfect agreement between all raters, and lower values are interpreted on a scale similar to that assumed for the Cohen Kappa. For a scheme with n raters and K scoring categories, Fleiss (Fleiss, 1971) calculates the proportion p_j of all assignments, for all raters and all subjects, to each category j, for j=1, 2, …, K, as follows:

$$p_j = \frac{1}{N \times n} \sum_{i=1}^{N} n_{ij} \tag{15}$$

where n_{ij} is the number of raters who assign subject i to category j. The proportion of rater pairs who agree for subject i can now be written (with capital P) as:

$$P_i = \frac{1}{L} \sum_{j=1}^{K} n_{ij} \times (n_{ij} - 1)/2 \tag{16}$$

where $L = n(n\text{-}1)/2$ which is the number pairs that are possible with n raters. The proportion of rater pairs that agree over all raters and all subjects is now:

$$P_O = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{17}$$

Fleiss then estimates the probability of agreement 'by chance' as:

$$P_e = \sum_{j=1}^{K} p_j^{\,2} \tag{18}$$

Substituting these values of P_o and P_e into the Kappa equation (3) gives an expression for Fleiss Kappa (Fleiss, 2011) which is widely used (Banerjee et al., 1999).

This expression does not generalise the unweighted Cohen Kappa as many researchers believe. The reason is that equation (18) makes the assumed distribution of 'by chance' scores the same for all raters. For all raters, the probability of getting score j by chance is assumed to be p_j as defined by equation (15). The definition of 'agreement by chance' is therefore different from that underlying the Cohen Kappa.

5. Multi-Rater versions of the Cohen Kappa

As explained by Warrens (Warrens, 2010), multi-rater versions of the Cohen Kappa have been proposed by Light (Light, 1971) and Hubert (Hubert, 1977) for categorical scoring. Conger

(Conger, 1980) generalises the formulation by Light (Light, 1971) and compares it with those by Fleiss (Fleiss, 1971) and Hubert (Hubert, 1977). According to Warrens (Warrens, 2010), the version by Hubert redefines P_e in terms of all possible pairs (r,s) of raters with r not equal to s, as follows:

$$P_e = \sum_{j=1}^{K} \frac{1}{L} \sum_{r=1}^{n} \sum_{s=r+1}^{n} p(r, j) \times p(s, j) \tag{19}$$

where $p(r, j)$ for $j = 1, 2, \ldots, K$ is the proportion of the N subjects that rater r assigns to scoring category j. Similarly, $p(s, j)$ is the proportion that rater s assigns to scoring category j. With this re-formulation for P_e, and P_o defined as for the Fleiss Kappa, equation (3) becomes a generalisation of the Cohen Kappa in the sense that:

$$P_o = \frac{1}{L} \sum_{r=1}^{n} \sum_{s=r+1}^{n} P_o(r, s) \quad \text{and} \quad P_e = \frac{1}{L} \sum_{r=1}^{n} \sum_{s=r+1}^{n} P_e(r, s) \tag{20}$$

where $P_o(r,s)$ and $P_e(r,s)$ are the P_o and P_e terms in equation (3) that define the Cohen Kappa between raters r and s. With the number of raters n equal to 2, this formulation becomes precisely the unweighted Cohen Kappa.

The alternative version by Light (Light, 1971) takes the multi-rater Cohen Kappa to be the arithmetic mean of the pair-wise Cohen Kappas for all possible pairs of raters. Clearly this multi-rater version also generalises the pair-wise Cohen Kappa, but it is, in general, different from both the Hubert (Conger) version and the Fleiss Kappa. Where the distribution of scores is the same for all raters, the Hubert (Hubert, 1977), Light (Light, 1971) and Fleiss (Fleiss, 1971) Kappas will all be identical. Where the distributions are not too dissimilar, as will often be the case, these three versions will be fairly close, though not identical.

Both versions of the multi-rater Cohen Kappa differ from the Fleiss formulation (Fleiss, 1971) because Fleiss specifies that each rater index does not necessarily refer to the same person. According to Fleiss (Fleiss, 1971), 'rater r' refers to a 'rater seat' rather than a specific scoring person. In this case, it is considered inappropriate to define P_e in terms of the distribution of scores at each seat. The more general assumption made by Fleiss about the likely distribution of scores at each seat is then more appropriate. The original Fleiss Kappa (Fleiss, 1971) is unaffected by the characteristic trends in the scoring by individuals. Only the distribution of scores among the K scoring categories is considered important. This is not the case with pair-wise Cohen Kappa.

The differences between the Fleiss Kappa and both versions of the 'multi-rater Cohen Kappa mentioned here are often small but noticeable. Where there are individual (fixed) raters rather than 'rater seats' it is probably best to use a multi-rater Cohen Kappa rather than the Fleiss Kappa (Fleiss, 1971) for multi-rater assessment. This is because it preserves the original definition of 'agreement by chance', which takes into account the (assumed) typical scoring distributions of the raters.

6. Multi-rater Consistency Measures Generalised to Ordinal Scoring

So far, multi-rater Cohen Kappa has only been defined for categorical scoring and it is widely stated that the Fleiss Kappa is only applicable to categorical scoring. The idea of expressing the pair-wise Cohen Kappa in terms of disagreement rather than agreement led to the development of weighted Cohen Kappa. There is no reason why the same approach should not be used for a multi-rater Cohen Kappa to obtain a weighted version. Generalising equation (4) to n raters, affording L = n(n-1)/2 rater pairs, gives the following equation for the Hubert (Hubert, 1977) multi-rater Cohen Kappa:

$$CK = 1 - \frac{\sum_{r=1}^{n}\sum_{s=r+1}^{n} D_o(r,s)}{\sum_{r=1}^{n}\sum_{s=r+1}^{n} D_e(r,s)} \tag{21}$$

where generalising equations (5) and (6) gives:

$$D_o(r,s) = \frac{1}{N}\sum_{i=1}^{N} C(A(i,r), A(i,s)) \tag{22}$$

$$D_e(\mathrm{r},\mathrm{s}) = \frac{1}{N^2}\sum_{j=1}^{K}\sum_{k=1}^{K} A_{jr}A_{ks}C(\alpha(j),\alpha(k)) \tag{23}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i,r), A(j,s)) \tag{24}$$

and A_{jr} and A_{ks} are the number of subjects which raters r and s score as scoring categories $\alpha(j)$ and $\alpha(k)$ respectively. With cost function C defined as 'unweighted' by equation (7), equation (21) becomes precisely the multi-rater Cohen Kappa previously defined in terms of agreement rather than disagreement. However, with C defined as 'linearly weighted' by equation (8), equation (21) becomes a 'linearly weighted' multi-rater Cohen Kappa, now

applicable to ordinal data. With C defined as 'quadratically weighted' by equation (9), equation (21) becomes a quadratically weighted multi-rater Cohen Kappa.

With n raters scoring N subjects to obtain scores $\{A(i,r)\}_{1,N}$ for $r = 1, 2, \ldots, n$, the multi-rater weighted Cohen Kappa formula can be written:

$$wCK = 1 - \frac{(1/N)\sum_{r=1}^{n}\sum_{s=r+1}^{n}\sum_{i=1}^{N} C(A(i,r)\ ,\ A(i,s))}{(1/N^2)\sum_{r=1}^{n}\sum_{s=r+1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i,r)\ ,\ A(j,s))} \tag{25}$$

This takes into account all possible differences between all possible pairs of different raters. It extends the linearly weighted multi-rater Kappa derived by Jalalinajafabadi (Jalalinajafabadi, 2016). A corresponding but different expression may be obtained by generalizing the Light (Light, 1971) version of the multi-rater Cohen Kappa to ordinal scoring with weighting.

The original Fleiss Kappa (Fleiss, 1971), as defined for non-fixed raters, can also be generalised to ordinal scoring with linear, quadratic or other weighting. The generalised version of Fleiss Kappa is given by equation (26):

$$wFK = 1 - \frac{\frac{1}{NL}\sum_{r=1}^{n}\sum_{s=r+1}^{n}\sum_{i=1}^{N} C(A(i,r)\ ,\ A(i,s))}{\frac{1}{(Nn)^2}\sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i,r)\ ,\ A(j,s))} \tag{26}$$

where the cost function C may apply no weighting, or linear, quadratic or other weighting This equation is valid for any number, n, of non-fixed raters including n = 2.

7. Multi-Rater Version of ICC

The multi-rater version of *ICC* (Müller & Büttner, 1994) generalises the pair-wise version defined by equation (2), to accommodate n raters as follows:

$$ICC = \frac{\frac{1}{L}\sum_{i=1}^{N}\sum_{r=1}^{n}\sum_{s=r+1}^{n}(A(i,r)-m)(A(i,s)-m)}{\frac{1}{n}\sum_{i=1}^{N}\sum_{r=1}^{n}(A(i,r)-m)^2} \tag{27}$$

$$= \frac{(1/(NL))\sum_{i=1}^{N}\sum_{r=1}^{n}\sum_{s=r+1}^{n}A(i,r)A(i,s) - m^2}{(1/(Nn))\sum_{i=1}^{N}\sum_{r=1}^{n}A(i,r)^2 - m^2} \quad (28)$$

$$\text{where} \qquad m = \frac{1}{nN}\sum_{r=1}^{n}\sum_{i=1}^{N}A(i,r) \quad (29)$$

Therefore, m is the pooled arithmetic mean of scores over all subjects and all raters. As demonstrated by equations (12) and (13), quadratically weighted Cohen Kappa and pair-wise *ICC* will be approximately the same when the pair-wise geometric and arithmetic means of the scores for each rater are approximately the same. It follows that quadratically weighted multi-rater Cohen Kappa and multi-rater *ICC* will be approximately the same when all pair-wise geometric and arithmetic means are approximately the same, as will often be the case. They are not necessarily identical. The generalisation used by Light (Light, 1971) is not the same as that used by ICC (Müller & Büttner, 1994), therefore the differences between quadratically weighted multi-rater Cohen Kappa and multi-rater ICC will generally be greater if Light's formulation is adopted.. However, it may be shown that with quadratic weighting, equation (26) for weighted Fleiss Kappa becomes identical to ICC as defined by equation (28) for any number of raters including n=2. Therefore ICC, like the Fleiss Kappa disregards differences in the scoring patterns of individual raters.

8. Examples

Table 1 is reproduced as an example of how the scoring of ten subjects by 14 raters may be distributed among five scoring categories. From this table, it is possible to calculate a value of the original Fleiss Kappa' (Fleiss, 1971) using equations (15) to (18) and (3) above. The value obtained is 0.2099. It is not possible to calculate a value of multi-rater Cohen Kappa from Table 1 because the scoring by each rater must be known. Many possible distributions of rater scoring correspond to Table 1 and it is straightforward to generate random examples of these. Table 2 is just one example. For the distribution of rater scoring shown in Table 2, the Hubert (Hubert, 1977) multi-rater Cohen Kappa, calculated from equations (15) to (17), (19) and (3), is found to be 0.2210 which differs from the original Fleiss Kappa value by about 5.29%. The Light (Light, 1971) value, obtained by averaging the values of pair-wise Cohen Kappa for all 91 possible pairs of raters, is 0.2263 which differs from the Fleiss Kappa by about 7.80 %. Further

examples may be generated with different values of multi-rater Cohen Kappa which may be closer to or further away from the original Fleiss Kappa.

It is appropriate to use multi-rater Cohen Kappa in cases where a fixed group of raters is engaged. It would not be appropriate for the example used by Fleiss (Fleiss, 1971) with non-fixed raters. Also, it would not be appropriate to calculate *ICC* or weighted versions of the Fleiss Kappa where the scoring is categorical. The AC_1 coefficient by Gwet (Gwet, 2014) for the scoring in Table 2 is 0.2256 which is close to the multi-rater Light and Hubert coefficients.

The effect of weighting may be illustrated by assuming that the scoring in Table 2 is ordinal. In this case the linear weighted Hubert, Light and Fleiss versions of Kappa are: 0.3944, 0.3975 and 0.3929 respectively. The quadratically weighted versions are: 0.5335, 0.5384 and 0.5405. As expected, the 5-rater ICC value is exactly the same as the quadratically weighted Fleiss Kappa value. The Hubert, Light and Fleiss Kappas and AC_1 are fairly close for this example because the distribution of scoring is similar for each rater. Other examples can produce greater differences.

Bearing in mind the suggestion that a form of the Cohen Kappa may be useful as a supplement or replacement for measures of CVI (Polit & Beck (2006), we calculated the unweighted Cohen Kappa for the questionnaire listed in Table 3.1 of (Sukron & Phutthikhamin, 2016) as assessed by five experts. The experts were the raters and the subjects became the 26 questions being assessed for their content validity. There is strong agreement among the experts and a high I-CVI index of 0.92 was unsurprisingly obtained. However, the multi-rater Hubert (Conger), Light and Fleiss Kappas produced values of 0.0055, -0.00618 and -0.0152 respectively. These values, suggesting that the consistency between experts is 'slight', appear misleading for this example. The AC_1 coefficient of Gwet (Gwet, 2014) was 0.7093 which is much more reasonable as a measure of consistency. This result provides a further example of the 'Kappa paradox' studied by Gwet (Gwet, 2014) and reason to adopt the AC_1 coefficient instead of the Cohen or Fleiss Kappa in applications where there is high agreement between raters or experts. AC_1 may be weighted (Gwet, 2014), like the Cohen and Fleiss Kappas, and it will be useful to explore the application of weighted consistency measurements in content validity assessments.

9. Conclusions

Expressions for the pair-wise Cohen Kappa are derived and generalized to two different multi-rater versions which are compared with the Fleiss Kappa. Both multi-rater Cohen Kappas

and the Fleiss Kappa may be weighted and applied to ordinal as well as categorical scoring. The conditions under which quadratically weighted pair-wise and multi-rater Cohen Kappa are equivalent to, respectively, pair-wise and multi-rater ICC have been clarified. A well known scoring example is presented to highlight the discrepancy between the original Fleiss Kappa (Fleiss, 1971) and the multi-rater Cohen Kappas. Where fixed raters are engaged, it may be considered more appropriate to use a multi-rater Cohen Kappa for measuring consistency rather than the original Fleiss Kappa since the scoring patterns of individual raters are taken into account. However the original Fleiss version remains the most appropriate Kappa for non-fixed raters. The AC_1 measure of consistency by Gwet is less well known, but may be preferable in view of its performance when there is high agreement among raters. The relationship between Cohen and Fleiss Kappa measurements of rater agreement and CVI measurements of content validity suggest that either Kappa may be useful as a supplement to CVI measures. However, as illustrated by example in Section 8, AC_1 may be preferred for CVI measurements where there is often high agreement among the experts and the Kappa paradox may be observed.

10. Acknowledgment

We acknowledge the work of Dr. Farideh Jalalinajafabadi and inspiration and encouragement from other colleagues, including Dr. Gavin Brown and Prof. Mikel Lujan, in the School of Computer Science, University of Manchester.

Table 1: *Scoring category distribution table (reproduced as an example)*

	Categories				
Subject	**1**	**2**	**3**	**4**	**5**
1	0	0	0	0	14
2	0	2	6	4	2
3	0	0	3	5	6
4	0	3	9	2	0
5	2	2	8	1	1
6	7	7	0	0	0
7	3	2	6	3	0

8	2	5	3	2	2
9	6	5	2	1	0
10	0	2	2	3	7

Table 2: *Example of Rater Scores corresponding to Table 1*

	Raters													
Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	5	5	5	5	5	5	5	5	5	5	5	5	5	5
2	4	3	4	3	3	3	5	2	2	3	4	3	4	5
3	3	3	5	4	5	4	5	5	4	5	3	5	4	4
4	3	3	4	4	3	3	3	2	2	3	2	3	3	3
5	3	3	4	3	1	3	1	3	2	3	2	5	3	3
6	2	2	1	2	2	1	1	1	2	1	1	2	1	2
7	4	3	4	3	1	3	1	3	2	3	3	2	4	1
8	5	3	4	3	5	4	1	2	2	2	2	2	1	3
9	2	3	1	2	3	1	1	4	2	1	1	2	1	2
10	5	3	4	3	5	5	5	2	2	5	5	5	4	4

REFERENCES

Banerjee M., Capozzoli M., McSweeney L. & Sinha D. (1999). Beyond Kappa: A Review of Interrater agreement measures, The Canadian Journal of Statistics, 27 (1), 3-23. https://doi.org/10.2307/3315487

Cohen J. (1960). A coefficient of agreement for nominal scales. Educational and Psychosocial Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

Cohen J. (1968). Weighted Kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213. https://doi.org/10.1037/h0026256

Conger A.J. (1980). Integration and Generalisation of Kappas for Multiple Raters. Psychol Bull., 88, 322-328. https://doi.org/10.1037/0033-2909.88.2.322

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

Fleiss J.L. & Cohen J. (1973). The equivalence of weighted Kappa and the intra class correlation coefficient as measures of reliability. J Educational and Psychological Measurement, 33, 613-619. https://doi.org/10.1177/001316447303300309

Fleiss J.L. (2011). Design and analysis of clinical experiments, John Wiley & Sons.

Gwet K. L. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.

Hubert l. (1977). Kappa Revisited, Psychol Bull, 84, 289-297. https://doi.org/10.1037/0033-2909.84.2.289

Jalalinajafabadi F. (2016). Computerised assessment of voice quality , Phd Thesis. University of Manchester.

Kitreerawutiwo, N. & Mekrungrongwong S. (2015). Health Behavior and Health Need Assessment among Elderly in Rural Community of Thailand: Sequential explanatory mixed methods study. LIFE: International Journal of Health and Life-Sciences, 1 (2), 62-69 https://doi.org/10.20319/lijhls.2015.12.6269

Koch G.G. (1982). Intraclass correlation coefficient. *Encyclopedia of Statistical Sciences*.

Light R.J. (1971). Measures of response agreement for qualitative data: some generalisations and alternatives. Psychol Bull, 76, 365-377. https://doi.org/10.1037/h0031643

Lee Rodgers J. & Nicewander W.A. (1998). Thirteen ways to look at the correlation coefficient. The American Statistician. 42(1), 59-66. https://doi.org/10.1080/00031305.1988.10475524

Müller R. & Büttner P.A. (1994). A critical discussion of intraclass correlation coefficients. Statistics in Medicine, 13(23-24): 2465-76 https://doi.org/10.1002/sim.4780132310

Polit, D.F. & Beck C.T. (2006), The Content Validity Index, Are You Sure You Know What's Being Reported?: Critique and Recommendations, Research In Nursing & Health, 29, 489–497 https://doi.org/10.1002/nur.20147

Rödel E.(1971). Fisher R.A. Statistical Methods for Research Workers, 14. Aufl., Oliver & Boyd, Edinburgh, London. XIII, 362 S., 12 Abb., 74 Tab., 40 s. *Biometrical Journal*. 13(6), 429-30. https://doi.org/10.1002/bimj.19710130623

Sukron & Phutthikhamin N. (2016), The Development Of Caregivers' Knowledge About Stroke And Stroke Caregiving Skills Tools For Stroke Caregivers In Indonesia, LIFE: International Journal Of Health And Life-Sciences, 2 (2), 35-47

Viera A.J., & Garrett J.M.(2005). Understanding interobserver agreement: the Kappa statistic. Fam Med, 37(5), 360-363.

Warrens M.J. (2010). Inequalities between Multi-Rater Kappas, *Advances in Data Analysis & Classification*, 4 no. 4, 271- 286. https://doi.org/10.1007/s11634-010-0073-4