A Study of Chance-Corrected Agreement Coefficients for the Measurement of Multi-Rater Consistency

Zheng Xie
School of Engineering
University of Central Lancashire
Preston, UK
e-mail: 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, UK
e-mail: cgadepalli@gmail.com

Barry M.G. Cheetham
School of Computer Science
University of Manchester
Manchester, UK
e-mail: barry@manchester.ac.u

***Abstract* — Chance corrected agreement coefficients such as the Cohen and Fleiss Kappas are commonly used for the measurement of consistency in the decisions made by clinical observers or raters. However, the way that they estimate the probability of agreement (Pe) or cost of disagreement (De) 'by chance' has been strongly questioned, and alternatives have been proposed, such as the Aickin Alpha coefficient and the Gwet AC1 and AC2 coefficients. A well known paradox illustrates deficiencies of the Kappa coefficients which may be remedied by scaling Pe or De according to the uniformity of the scoring. The AC1 and AC2 coefficients result from the application of this scaling to the Brennan-Prediger coefficient which may be considered a simplified form of Kappa. This paper examines some commonly used multi-rater agreement coefficients including AC1 and AC2. It then proposes an alternative subject-by-subject scaling approach that may be applied to weighted and unweighted multi-rater Cohen and Fleiss Kappas and also Intra-Class Correlation (ICC) coefficients.**

Keywords - Rater consistency, Fleiss Kappa, Cohen Kappa, ICC, Gwet's AC1 coefficient

I. INTRODUCTION

Studies of the consistency of clinical assessments are important in medical research. A patient will often be observed by one clinician, at least initially. The patient may be reassured to learn that the resulting assessment is likely to be independent of the choice of clinician since it is likely to affect any prescribed treatment. If the assessments of certain conditions are often found to be different for different clinicians, this may not necessarily be a bad thing as different perspectives can be valuable. However, it would be useful to know to what extent the assessments are likely to be consistent and whether certain observations are hard or easy to assess. Such knowledge may, for example, suggest when a second opinion may be valuable.

Investigating how much consistency is likely, and finding ways of improving it requires a clinical trial with a selection of subjects comprising patients and other volunteers and a number of clinical observers referred to as raters. Such a trial was carried out by Gadepalli et al [1] for a voice quality assessment procedure with 102 subjects and five raters. This trial required measurements of the intra-rater (self) consistency of decisions by the same raters at different times and also of inter-rater consistency of decisions by different raters observing the same subjects.

The decisions may be diagnoses of medical conditions or the severity of such conditions. The decisions may be categorical and denoted by labels. Or they may by ordinal which means that they are numbers often referred to as scores. Unlike labels, scores have magnitudes and may be compared in terms of differences between them.

Given N subjects and R raters who each observe all subjects, the 'proportion of agreement' Po may be considered as a measure of consistency for categorical or ordinal decisions [2]. Denoting by A(i,r) the decision or score given by rater r to subject i, Po may be expressed as:

$$P_o = \frac{1}{NL}\sum_{i=1}^{N}\sum_{r=1}^{R}\sum_{s=r+1}^{R}\delta(A(i,r),A(i,s)) \text{ where } \delta(u,v)=\begin{cases}1: u=v\\ 0: u\neq v\end{cases} \quad (1)$$

In equation (1), $L = R(R\text{-}1)/2$ which is the number ways in which two distinct raters, r and s, may be selected from the R raters for comparison of their decisions. Therefore P_o counts the number of times that a pair of raters agree for a subject. It is expressed as a proportion of the number of subjects times the number of rater pairs, and is a number between 0 and 1. When there is complete agreement by all raters for all subjects, P_o will be equal to 1. When there is almost no agreement, P_o will be close to zero.

Although equation (1) is intended for categorical (nominal) decisions, it may be used also for ordinal scoring if the scores are considered as labels rather than numbers. When used for ordinal decisions (scores), equation (1) gives equal weight to all possible differences regardless of their magnitudes. However, for ordinal scoring it is often preferable to give more importance to larger differences than smaller difference, and this leads to a weighted version of P_o defined as follows:

$$P_o = \frac{1}{NL} \sum_{i=1}^{N} \sum_{r=1}^{R} \sum_{s=r+1}^{R} w(A(i,r), A(i,s)) \qquad (2)$$

where w(u,v) is a 'weighting function' [3], which may be expressed in terms of a 'cost function' $C(u,v)$ as follows:

$$w(u,v) = 1 - C(u,v) \qquad (3)$$

Assuming that there are Q possible scores 1, 2, …, Q, for linear weighting:

$$C(u,v) = |u - v| / (Q-1) \qquad (4)$$

for quadratic weighting,

$$C(u,v) = (u-v)^2 / (Q-1)^2 \qquad (5)$$

and for no weighting,

$$C(u,v) = 1 - \delta(u,v) \qquad (6)$$

When $C(u,v)$ is defined by equation (6), equation (2) becomes identical to equation (1) as applied to ordinal scores considered as labels. There are many other possible cost-functions that may be considered, but these three are of special interest. In equations (4) to (6), $C(u,v)$ is the cost of any disagreement between score u and score v. The cost determines the degree to which the value of P_o is decreased from unity by a rater-pair disagreement. In all cases, the cost of the maximum possible disagreement is 1 and the cost is zero when there is no disagreement. With linear weighting, $C(u,v)$ is proportional to the magnitude of the score difference, with quadratic weighting, this magnitude is squared and with no weighting, any score difference contributes the same unit cost. As with the unweighted version of P_o, the weighted version is equal to 1 for perfect agreement, and the minimum possible value is equal to 0. A value of P_o close to zero would indicate that all rater-pairs disagree to the maximum possible extent for all subjects.

P_o may be re-expressed as:

$$P_o = 1 - D_o \qquad (7)$$

where D_o is the overall unweighted or weighted cost of disagreement defined as:

$$D_o = \frac{1}{NL} \sum_{i=1}^{N} \sum_{r=1}^{R} \sum_{s=r+1}^{R} C(A(i,r), A(i,s)) \qquad (8)$$

II. Chance-Corrected Agreement Coefficients

Unweighted and weighted versions of P_o are straightforward measures of consistency for categorical or ordinal scoring. But they are biased by the probability of some agreement occurring by chance. If all raters were to make random decisions evenly distributed over Q categories or scores, 'by chance' agreement would be expected with a probability of $1/Q$, even if the raters made their decisions without reference to the subjects. This would make unweighted P_o equal to 1/Q and weighted P_o equal to T_w/Q^2 where [3]:

$$T_w = \sum_{k=1}^{Q} \sum_{\ell=1}^{Q} w(k,\ell) \qquad (9)$$

With four scoring categories and 'by chance' scoring, the expectation of unweighted P_o would be 1/4 or 25% for an even spread of decisions over the four categories, and with an uneven spread of decisions, P_o could be even greater, thus giving a false impression of some consistency when there may be none.

Chance corrected agreement coefficients aim to cancel out the bias in P_o, while still providing a number between 0 and 1. They are normally expressed as:

$$\gamma = \frac{P_0 - P_e}{1 - P_e} \qquad (10)$$

where, for categorical or unweighted scoring, P_o is as defined by equation (1) and P_e is an estimate of the probability of agreement by chance. P_e may be re-expressed as:

$$P_e = 1 - D_e \qquad (11)$$

where D_e is an estimate of the probability of disagreement by chance.

To extend equations (10) and (11) to weighted ordinal scoring, P_o is generalised by equation (2) and D_e is generalised to be an estimate of the overall weighted cost of disagreement by chance. If there is almost complete agreement, P_o will be close to 1 and γ will be close to 1 unless P_e is also close to 1. If P_e is close to 1, almost all agreement or disagreement would be considered to have occurred by chance.

III. Brennan-Prediger Coefficient

The simplest chance corrected agreement coefficient is known as the Brennan-Prediger coefficient [4]. The unweighted or categorical version is:

$$\gamma = \frac{P_o - P_e}{1 - P_e} \quad \text{where} \quad Pe = \frac{1}{Q} \tag{12}$$

with P_o defined by equation (1). The weighted version for ordinal scoring is:

$$\gamma = \frac{P_o - P_e}{1 - P_e} \quad \text{where} \quad P_e = \frac{T_w}{Q^2} \tag{13}$$

with P_o defined weighted by equation (2) and T_w by equation (9). The Brennan-Prediger coefficient cancels out the bias in P_o when the rater decisions or scores are evenly distributed among the Q categories. But it will not do this accurately for uneven distributions.

IV. Multi-Rater Cohen Kappa

The Cohen Kappa aims to remove the bias present in P_o by estimating and taking into account the probability of agreement 'by chance' given the distribution of decisions produced by each rater. It was originally proposed [5] for categorical rating by two raters, and was generalised by Hubert [6] and Conger [7] to a multi-rater version that may be expressed [2] as follows:

$$\gamma = \frac{P_o - P_e}{1 - P_e} \quad \text{with} \quad P_e = \frac{1}{LN^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{r=1}^{R} \sum_{s=r+1}^{R} \delta(A(i,r), A(j,s)) \tag{14}$$

P_o is as defined by equation (1). As defined by equation (14), P_e is an estimate of the probability of a pair of arbitrary raters agreeing by chance when these raters make arbitrary decisions which are not necessarily the same as those of the actual raters, but are similarly distributed over the Q possible categories or scores. The distributions depend on the individual scoring characteristics of each rater and also on the diversity of the N subjects. Assuming that the scoring is applied to the severity of some condition such as voice quality impairment, ranging from normal to severe, if the diversity of the subjects is such that the distribution of degrees of severity across this range is fairly uniform, then the scoring will be expected to be fairly uniform. P_e will then be mainly dependent on the scoring characteristics of each rater and how these differ from rater to rater. If, however, there is an inherent bias towards a particular degree of severity, for example, a high number of severe cases, this will affect the rater scoring distributions in a way that is unrelated to the scoring characteristics of the raters.

For P_e to be a reliable estimate of 'by chance' scoring for a population of subjects, it has to be assumed that the N subjects are a reasonable sample of the population. If there is a bias in the scores obtained for the N subjects, then it is assumed that that the same bias exists in the population. Otherwise the sample will be unrepresentative of the population and the estimate of P_e will be unreliable.

Equation (14) becomes identical to the original Cohen Kappa when the number of raters, R, is equal to two. An alternative generalisation by Light [8] is also identical for two raters but slightly different for more. Equation (14) may be further generalised to weighted form [9] for ordinal scoring by defining P_o by equation (2) and P_e by equation (15):

$$P_e = 1 - D_e = 1 - \frac{1}{LN^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{r=1}^{R} \sum_{s=r+1}^{R} C(A(i,r), A(j,s)) \tag{15}$$

As with the Brennan-Prediger coefficient, the introduction of weighting increases the cost of any by chance disagreement between arbitrary raters. But P_e is no longer a constant that is independent of the distribution of rater scores. Combining equations (2), (10) and (15) we obtain an expression for the weighted multi-rater Cohen Kappa which can be expressed [2] as follows:

$$\gamma = 1 - \frac{(1/N) \sum_{r=1}^{R} \sum_{s=r+1}^{R} \sum_{i=1}^{N} C(A(i,r)\ ,\ A(i,s))}{(1/N^2) \sum_{r=1}^{R} \sum_{s=r+1}^{R} \sum_{i=1}^{N} \sum_{j=1}^{N} C(A(i,r)\ ,\ A(j,s))} \tag{16}$$

V. Fleiss-Kappa

The Fleiss Kappa [10] is an alternative chance-corrected agreement coefficient originally defined for two or more categorical raters. When used for two raters, it becomes identical to the Scott Pi coefficient [11]. 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 categorical scoring with R raters and Q scoring categories, Fleiss [10] calculates P_o by equation (1) and then calculates the proportion, π_k, of all assignments to category k, for all raters and all subjects, for k = 1, 2, …, Q, as follows:

$$\pi_k = \frac{1}{NR} \sum_{i=1}^{N} \sum_{r=1}^{R} \delta(A(i,s), k)) \tag{17}$$

If it is assumed that the N subjects are a reasonable sample of the population, π_k may be considered an estimate of the probability that a randomly selected rater will classify a randomly selected subject into category k. As there are Q categories, it follows that:

$$\sum_{k=1}^{Q} \pi_k = 1 \tag{18}$$

Fleiss [10] then estimates the probability of agreement 'by chance' as:

$$P_e = \sum_{k=1}^{Q} {\pi_k}^2 \tag{19}$$

A weighted version [3] of the Fleiss Kappa is obtained by defining P_o by equation (2) and P_e by equation (20):

$$P_e = 1 - D_e \ = 1 - \sum_{k=1}^{Q}\sum_{\ell=1}^{Q} C(k,\ell)\pi_k \pi_\ell \tag{20}$$

Substituting these values of P_o and P_e into equation (10) gives an expression for Fleiss Kappa which can be re-expressed [2] as:

$$\gamma \ = \ 1 \ - \ \frac{\dfrac{1}{NL}\displaystyle\sum_{r=1}^{R}\sum_{s=r+1}^{R}\sum_{i=1}^{N} C(A(i,r)\ ,\ A(i,s))}{\dfrac{1}{(Nn)^2}\displaystyle\sum_{r=1}^{R}\sum_{s=1}^{R}\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i,r)\ ,\ A(j,s))} \tag{21}$$

The cost function C may apply no weighting, or linear, quadratic or other weighting as in equations (4 - 6). Equation (21) is valid for any number, R, of raters including $R = 2$.

The Fleiss and Cohen versions of Kappa differ because Fleiss specifies that each rater index does not necessarily refer to the same person. In this case, it is considered inappropriate to define P_e in terms of the distribution of scores for each rater index. The more general assumption made by Fleiss about the likely distribution of scores for each rater index, i.e. that they are all equal, is more appropriate. The Fleiss Kappa [10] is unaffected by the characteristic trends in the scoring by individuals. Only the distribution of scores among the Q scoring categories by all raters is considered important, though, as for the Cohen Kappa, this is affected by any inherent bias in the diversity of the N subjects. The differences between the Fleiss Kappa and the multi-rater Cohen Kappa are often small but sometimes noticeable. Where there are individual (fixed) raters it is appropriate to use the multi-rater Cohen Kappa rather than the Fleiss Kappa because it takes into account the typical scoring characteristics of the individual raters.

VI. Missing scores

The equations given above for the multi-rater Cohen, Fleiss and Brennan-Prediger coefficients assume that all N subjects are scored by all R raters. They have been generalised by Gwet [3] to the case where some scores are missing, but we do not consider this case here.

VII. Gwet's Paradox

There is controversy about the way the Cohen Kappa [5] and Fleiss Kappa [10] estimate by chance agreement (14). Different approaches, such as the AC_1 and AC_2 coefficients by Gwet [3], are gaining currency. The deficiencies of the Cohen & Fleiss Kappas are illustrated by the following example which is similar to examples quoted by Gwet [3].

In this example, there are two raters for 20 subjects with two possible scoring categories. Rater 1 scores all subjects in category 1, and rater 2 scores 18 out of 20 in category 1. It may appear that that there is a high level of agreement. However, since $P_o = 0.9$ and $P_e = 0.9$ for unweighted Cohen Kappa and $P_e = 0.905$ for unweighted Fleiss Kappa, both Kappas give zero or a value close to zero, thus indicating little or no agreement. The problem lies with the estimation of P_e, since almost all agreement is classified as occurring by chance. This is because the scoring of the raters is assumed to be biased towards 1. If the diversity of the subjects is such that the anticipated scores are fairly evenly spread across the range 1 to Q (with Q=2), the rater scores would then be unrelated to the subjects so that any agreement would mostly be by chance. However, if the subjects themselves are inherently biased towards the score of 1, this could account for the scores in Table 1, making them highly consistent and the Kappa values misleading. By equation (12), the Brennan-Prediger coefficient, which is independent of the distribution of scores, gives the more intuitive value of 0.8.

Ideally, when defining a chance-corrected agreement measure we should specify the expected scoring characteristics for a population of subjects. If this is a population for which the overwhelming majority of subjects are expected to be scored as category 1, then the Cohen and Fleiss Kappas obtained from Table 1 are highly pessimistic. The Brennan-Prediger coefficient disregards the effect on P_e of the distribution of actual scores in Table 2 and assumes an even distribution. The Cohen and the Fleiss Kappa take the actual distribution into account, while assuming a subject group with an even distribution of degrees of severity. This assumption appears unlikely to be appropriate and is responsible for the paradox.

VIII. Effect of Distribution of Scoring

The The paradox illustrated above, and referred to by Gwet [3], occurs whenever the value of π_k becomes close to 1 for some value of k. In this case, by equation (18), all other values of π_k become close to zero. To investigate this situation, we randomly generated a set of scores for $N = 50$ subjects, $R = 5$ raters and $Q = 4$ scoring categories. Given Q values of π_k, we generated a random score in the range 1 to Q (inclusive) for each subject index i, for each of the R raters, such that the overall probability of getting score k was equal to π_k for $k = 1, 2, \ldots, Q$.

Initially, we made $\pi_k = 1/Q$ for $k = 1, 2, \ldots, Q$, which meant that all scores were equally probable over all subjects and all raters. By equations (14) and (19), this case gives $P_e = 1/Q$ for both the unweighted Cohen and Fleiss Kappas, which are lowest possible values. P_e is always equal to $1/Q$. for the unweighted Brennan-Prediger coefficient.

We then randomly generated further sets of random scores, with one of the π_k values increased from $1/Q$, and the other $(Q\text{-}1)$ values decreased to satisfy equation (19). We chose π_1 to be the value that increased, and made all other π_k values equal to $(1\text{-} \pi_1)/(Q\text{-}1)$. By gradually increasing π_1 towards 1 we generated a series of scoring patterns that gradually approached the maximally concentrated distribution where $\pi_1 = 1$ and all other values of π_k are zero. For this maximally concentrated distribution, all raters give score1, to all subjects.

The situation when π_1 becomes close to 1 further demonstrates the paradox pointed out by Gwet [3]. The resulting values of unweighted Cohen and Fleiss Kappas are plotted against increasing π_1 in Figure 1, along with the Brennan-Prediger coefficient and the Gwet AC_1 coefficient. It may be seen in Figure 1 that as π_1 approaches 1, both the Cohen and Fleiss Kappas remain close to zero indicating no agreement except by chance. The Brennan Prediger coefficient approaches 1 (perfect agreement) as π_1 approaches 1.

The corresponding values of P_e for each of the coefficients plotted in Figure 1 are plotted against π_1 in Figure 2. Perhaps unexpectedly, the probability of agreement by chance, as estimated by both the unweighted Cohen and Fleiss Kappas, increases as the scores become more and more concentrated on score 1. P_e for the unweighted Brennan-Prediger coefficient remains constant at 1/Q with Q=4. The Gwet AC_1 coefficient will be discussed in the next Section.

IX. GWET'S AC_1 AND AC_2 COEFFICIENTS

Gwet's AC_1 and AC_2 coefficients can be considered generalisations of the unweighted and weighted Brennan-Prediger coefficients and they use the same value of unweighted or weighted P_o. In order to calculate a value of P_e (probability of agreement by chance) Gwet [3] adapts the idea used previously by Aickin [12] of dividing the subjects into those which are 'hard to score' and those which are 'easy to score', and estimating P_e for the 'hard' subjects only. It is suggested that the 'easy' subjects may be disregarded on the grounds that any agreement for easy subjects will not be by chance.

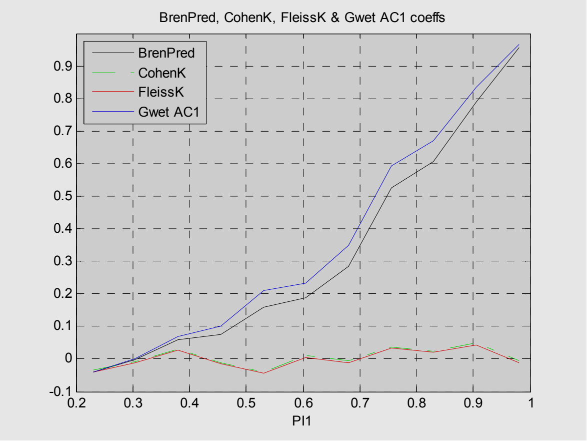

Figure 1. Graph comparing unweighted Cohen and Fleiss Kappas with the Brennan-Prediger and Gwet's AC_1 coefficients for increasing concentration of scores (Kappas almost coincide).

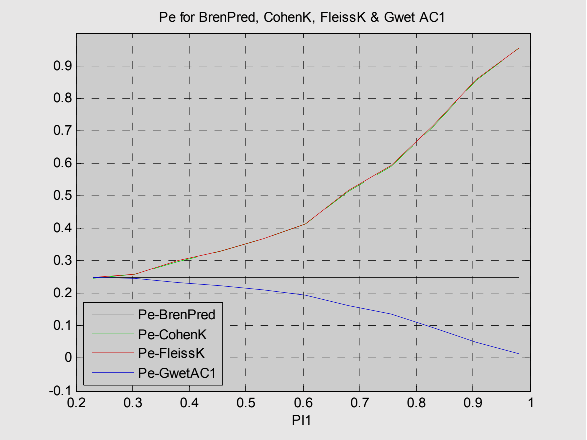

Figure 2: Graph of P_e against π_1 for unweighted Cohen and Fleiss Kappas and the Brennan-Prediger and Gwet AC_1 coefficients.

Gwet implements this idea in a probabilistic way by defining a function *P(R)* as the 'degree of uniformity' of the distribution of subjects across the Q categories. *P(R)* is estimated from the distribution of the scores given to all subjects by the *R* raters. The degree to which all raters give the same or similar scores to the *N* subjects is presumed to determine the degree to which the subjects are easy to score. A group of subjects given rater scores that are more uniformly distributed over the Q available scores is presumed to be harder to score, since there is less agreement in the scoring. *P(R)* lies between 0 and 1, and becomes close to 1 when the distribution of scores is close to being uniform, i.e. evenly spread out among all possible categories or scores. P(R) becomes close to zero when there is a strong bias towards one particular category or score as in Table 1. Gwet refers to *P(R)* as the probability of selecting a hard subject. Gwet's formula for *P(R)* is equation (22):

$$P(R) = \frac{\sum_{k=1}^{Q} \pi_k (1-\pi_k)}{1-1/Q} \tag{22}$$

The upper curve in Figure 3 shows *P(R)* plotted against π_1 for N=50, R=5 and Q=4. When $\pi_1 = 0.25$ the distribution of rater scores is even and all scores are equally likely. In this case, equation (22) gives a value of *P(R)* close to 1. As π_1 is increased towards 1, *P(R)* decreases towards zero.

The formulae for Gwet's AC$_1$ and AC$_2$ coefficients are obtained by modifying the Brennan-Prediger coefficient as follows:

$$AC_1 = \frac{P_o - P_e}{1 - P_e} \quad \text{where} \quad Pe = \frac{P(R)}{Q} \tag{23}$$

$$AC_2 = \frac{P_o - P_e}{1 - P_e} \quad \text{where} \quad P_e = \frac{1}{Q^2}\sum_{k=1}^{Q}\sum_{\ell=1}^{Q} w(k,l)\, P(R) \tag{24}$$

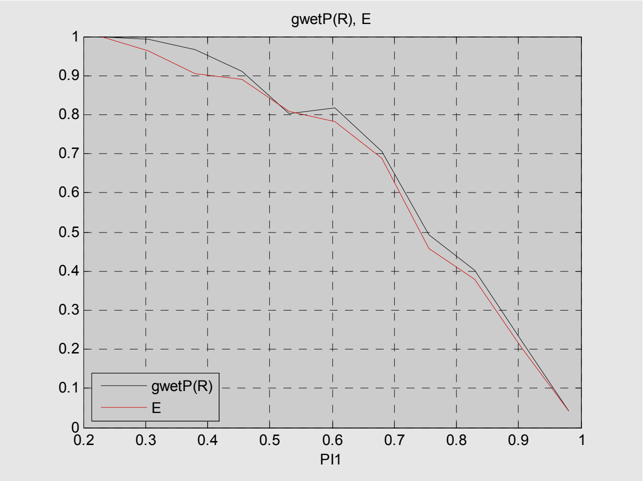

Figure 3. Gwet's P(R) function and function E

AC$_1$ applies to categorical or unweighted scoring and AC$_2$ to ordinal scoring with weighting. The effect of *P(R)* is to reduce the estimation of P_e as the degree of uniformity decreases. AC$_1$ is plotted against π_1 in Figure 1, and the corresponding values of P_e for AC$_1$ are plotted in Figure 2.

As may be expected, the behaviour of AC$_1$ is similar to that of the Brennan-Prediger coefficient. The effect of *P(R)* is to decrease P_e as the degree of concentration on a single score increases. This increases AC$_1$ in comparison to the Brennan-Prediger coefficient which continues to assume a uniform distribution of scores.

It seems reasonable to assume that P$_e$ should decrease as the distribution of scores decreases. In this case, the subjects are indeed likely to become easier to score. But there are other cases where subjects may become easier to score, but the scores are not concentrated on a single score.

Consider, for example, the scores given in Table 2. Almost all scores agree, therefore it must be inferred that these subjects are easy to score. But *P(R)* as defined by Gwet [3] will be close to 1 for this example. The description of *P(R)* as the probability of selecting a subject that is hard to score may therefore be misleading. It is better to describe *P(R)* in terms of the degree of uniformity of scoring. Despite the explanation given by Gwet [3], *P(R)* is really defined from the overall distribution of scores and not the hardness or easiness of scoring.

X. Application of *P*(*R*) Scaling to Cohen and Fleiss Kappas

Gwet [3] states that it would not be appropriate to take marginal probabilities (i.e. score distributions) into account when defining AC$_1$ and AC$_2$. Despite this assertion, there may be a case for applying a measure of the degree of uniformity to both the Cohen and Fleiss Kappas. Taking *P(R)* as such a measure, multiplying equations (15) and (20) for the Cohen and Fleiss Kappas respectively by *P(R)* gives the graphs referred to as CohenK-AC1 and FleissK-AC1 in Figure 4. The graphs almost coincide, and the paradox exhibited by the Cohen and Fleiss Kappas has now been eliminated.

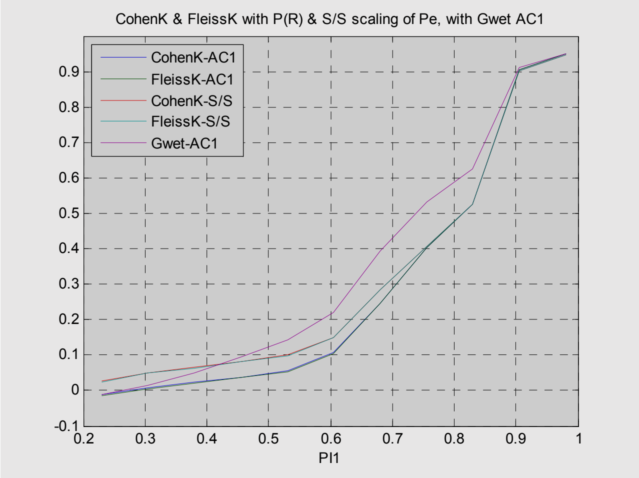

Figure 4. Cohen and Fleiss Kappas with P(R) and subject-by-subject scaling compared with Gwet's AC$_1$ coefficient.

XI. Subject-by-subject Implementation of Scaling

A more direct way of implementing the 'hard/easy to score' principle discussed by Gwet [3] is to apply it to each individual subject rather than applying it probabilistically. For each subject *i*, define a 'by chance' probability, *H(i)*, according to the rater scores it has been given and the overall spread of rater scores. Then the contribution to P$_e$ of any 'by chance' disagreement within rater pairs scoring subjects *i* and *j* may be scaled according to *H(i) and H(j)*. The maximum value of these probabilities determines the contribution.

For Cohen Kappa, P_e as previously defined by equation (15) becomes:

$$P_e = \frac{1}{LN^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{R}\sum_{s=r+1}^{R} w(A(i,r),A(j,s))\max(E(i),E(j)) \tag{25}$$

where *w(u,v)* = 1 − *C(u,v)* as in equation (3). The contribution to P_e from scores *A(i,r)* and *A(j,s)* are unaffected if either subject *i* or subject *j* is considered likely to have been scored by chance. If both subjects are considered less likely to have been scored by chance, the maximum of *H(i)* and *H(j)* will be reduced, perhaps to zero. Therefore, the contribution to P_e from scores *A(i,r)* and *A(j,s)* will be reduced, thus reducing the estimated probability of agreement by chance.

For the Brennan-Prediger coefficient, P_e in equation (13) becomes:

$$P_e = \frac{T_w}{Q^2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}}\max(\mathrm{E}(i),E(j)) \qquad (26)$$

which is valid for categorical or weighted ordinal scoring. For the Fleiss Kappa:

$$P_e = \frac{1}{R^2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{R}\sum_{s=r}^{R} w(A(i,r),A(j,s))\max(E(i),E(j)) \qquad (27)$$

This new approach requires a definition of 'by chance' probability *H(i)* for each subject.

There are several interesting ways to do this, but a simple and obvious one is to use equations (28) and (29) as suggested by equation (22) which defines *P(R)*.

$$\mathrm{H}(i) = \frac{\sum_{k=1}^{Q}\pi_k(i)(1-\pi_k(i))}{1-1/Q} \qquad (28)$$

$$\text{with} \quad \pi_k(i) = \frac{1}{R}\sum_{r=1}^{R}\delta(A(i,r),k)) \qquad (29)$$

Therefore *H(i)* quantifies the uniformity of the scores given by all *R* raters to a single subject *i* since $\pi_k(i)$ is the proportion of these *R* scores that is equal to *k*. Clearly,

$$\sum_{k=1}^{Q}\pi_k(i) = 1 \text{ for all } i \quad \text{and} \quad \pi_k = \frac{1}{N}\sum_{i=1}^{N}\pi_k(i) \qquad (30)$$

However, it must be noted that:

$$E = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} H(i) \neq P(R) \qquad (31)$$

Therefore we do not expect the subject-by-subject scaling by *H(i)* to be equivalent to the scaling by *P(R)* as defined by Gwet. Differences between this subject-by-subject approach and Gwet's *P(R)* implementation may be illustrated by comparing the curves in Figure 3. The black curve is *P(R)* and the red curve represents *E* as defined by equation (31). Both curves in Figure 3 reduce to zero as subjects become more and more concentrated on a single score. However *E,* when scaled to a maximum of 1, remains close to *P(R)* as π_1 approaches 1.

The probability measure defined by equation (28) was applied subject-by-subject to the unweighted Cohen and Fleiss Kappas by redefining P_e as in equations (25) and (27). The new forms of Kappa are referred to as CohenK-S/S and FleissK-S/S. The result is seen in Figure 4, and may be compared with the other measures. It may be seen that the Kappas almost coincide and are close to but generally lower than the AC_1 coefficient. The modification has eliminated the paradox that kept the Fleiss and Cohen Kappas close to zero in Figure 1. Applying the subject-by-subject implementation to the scores in Table 2 gives values of H(i) that are mostly zero. Therefore P_e will be small. This demonstrates that the hard-easy' principle of Gwet is now applied whether or not the majority of scores are concentrated on a small subset of scores.

XII. Intra-class Correlation (ICC)

Measurements of consistency in ordinal scoring are forms of correlation. The Pearson Correlation coefficient [13] is not normally appropriate for measuring consistency [14] as it takes into account only variations about the mean for each rater. However, the 'intra-class correlation' coefficient (*ICC*) [15] may be used as a consistency measure. The original form of ICC [16] for a pair of raters A and B may be written as follows:

$$ICC = \frac{\sum_{i=1}^{N}(A(i)-m)(B(i)-m)}{0.5\left(\sum_{i=1}^{N}(A(i)-m)^2 + \sum_{i=1}^{N}(B(i)-m)^2\right)} \qquad (32)$$

where N subjects are scored $\{A(i)\}_{1,N}$ by rater *A* and $\{B(i)\}_{1,N}$ by rater *B*, and *m* denotes the mean of the scores given by both *A* and *B*. Multi-rater ICC generalises the pair-wise version in equation (32) to *R* raters as follows:

$$ICC = \frac{\frac{1}{L}\sum_{i=1}^{N}\sum_{r=1}^{R}\sum_{s=r+1}^{R}(A(i,r)-m)(A(i,s)-m)}{\frac{1}{R}\sum_{i=1}^{N}\sum_{r=1}^{R}(A(i,r)-m)^2} \qquad (33)$$

$$= \frac{(1/(NL))\sum_{i=1}^{N}\sum_{r=1}^{R}\sum_{s=r+1}^{R}A(i,r)A(i,s) - m^2}{(1/(NR))\sum_{i=1}^{N}\sum_{r=1}^{R}A(i,r)^2 - m^2} \qquad (34)$$

$$\text{where} \quad \mathrm{m} = \frac{1}{\mathrm{RN}}\sum_{r=1}^{R}\sum_{i=1}^{N}A(i,r) \qquad (35)$$

with $A(i,r)$ denoting the score given by rater r to subject i and $L = R(R\text{-}1)/2$. Therefore, m is now the pooled arithmetic mean of scores over all subjects and all raters.

By equation (21) with $C(u,v)$ defined by equation (5), the quadratically weighted Fleiss Kappa for R raters is equal to:

$$\gamma = 1 - \frac{\frac{1}{NL}\sum_{r=1}^{R}\sum_{s=r+1}^{R}\sum_{i=1}^{N}(A(i,r) - A(i,s))^2}{\frac{1}{(NR)^2}\sum_{r=1}^{R}\sum_{s=1}^{R}\sum_{i=1}^{N}\sum_{j=1}^{N}(A(i,r) - A(j,s))^2} \tag{36}$$

$$= 1 - \frac{\frac{1}{NL}\sum_{r=1}^{R}\sum_{s=r+1}^{R}\sum_{i=1}^{N}(A(i,r)^2 + A(i,s)^2 - 2A(i,r)A(i,s))}{\frac{1}{(NR)^2}\sum_{r=1}^{R}\sum_{s=1}^{R}\sum_{i=1}^{N}\sum_{j=1}^{N}(A(i,r)^2 + A(j,s)^2 - 2A(i,r)A(j,s))} \tag{37}$$

$$= 1 - \frac{\frac{(R-1)}{NL}\sum_{r=1}^{R}\sum_{i=1}^{N}\left(A(i,r)^2\right) - \frac{2}{NL}\sum_{r=1}^{R}\sum_{s=r+1}^{R}\sum_{i=1}^{N}A(i,r)A(i,s)}{\frac{1}{(NR)^2}\sum_{r=1}^{R}\sum_{s=1}^{R}\sum_{i=1}^{N}\sum_{j=1}^{N}(2A(i,r)^2 - 2A(i,r)A(j,s))} \tag{38}$$

$$= 1 - \frac{\frac{2}{NR}\sum_{r=1}^{R}\sum_{i=1}^{N}\left(A(i,r)^2\right) - \frac{2}{NL}\sum_{r=1}^{R}\sum_{s=r+1}^{R}\sum_{i=1}^{N}A(i,r)A(i,s)}{\frac{2}{(NR)}\sum_{r=1}^{R}\sum_{i=1}^{N}(A(i,r)^2) - 2m^2} \tag{39}$$

$$= \frac{\frac{1}{NL}\sum_{r=1}^{R}\sum_{s=r+1}^{R}\sum_{i=1}^{N}\left(A(i,r)A(i,s)\right) - m^2}{\frac{1}{(NR)}\sum_{r=1}^{R}\sum_{i=1}^{N}(A(i,r)^2) - m^2} \tag{40}$$

As equation (40) is identical to equation (34), the quadratically weighted Fleiss Kappa for R raters is identical to the multi-rater ICC as defined above. It is well known that ICC and quadratically weighted Cohen Kappa are usually close, though not identical. However the identity established above may not be so widely known. Therefore, like the Fleiss Kappa and unlike the Cohen Kappa, ICC disregards differences in the scoring patterns of individual raters.

Since ICC is exactly equal to quadratically weighted Fleiss Kappa, it therefore may be considered to incorporate correction for chance agreement. Consequently, it produces the Gwet paradox when there is a concentration on one score. As with the Cohen and Fleiss Kappas, it may be modified by the application of Gwet's P(R) function or its subject-by-subject implementation.

XIII. Conclusions

There is a paradox with the Cohen and Fleiss Kappas which is observed when rater scores are concentrated on one score. Since Intra-classCorrelation (ICC) in its original form [16] and quadratically weighted Fleiss Kappa have been shown to be identical, the paradox will also occur with *ICC* when used for measuring rater agreement. It arises because the rater scores are used not only to quantify the actual agreement but also to estimate the probability of agreement by chance, or the cost of disagreement by chance (P_e). The basic problem is that P_e is inadequately defined by rater scores with a strong bias towards a small subset of the possible scores.

If it is assumed that all scores are equally likely in the population, the Brennan-Prediger coefficient correctly estimates P_e. As a means of catering for other distributions of scores, Gwet [3] sets out to improve this coefficient by de-emphasising the contributions to P_e from subjects that are considered easy to score. These subjects are considered unlikely to have been scored by chance. The de-emphasis is achieved by multiplying each contribution by a function *P(R)* which is a measure of the degree of uniformity in the scoring. The description of *P(R)* as the probability of selecting a subject which is hard to score is perhaps misleading. This is because there can be many subjects that may be considered easy to score but whose scores are not concentrated on one score or a small subset of scores. Nevertheless, there is a case for applying *P(R)* as a measure of uniformity to the Cohen and Fleiss Kappas and ICC to eliminate the paradox that may be exhibited by these coefficients.

We have investigated a subject-by-subject implementation of the Gwet 'hard-easy' principle with a 'by chance' probability estimated for each individual subject. It eliminates the paradox and has potential to improve the underlying estimate of the population statistics from the sample provided.

References

[1] Gadepalli C., Jalalinajafabadi F, Xie Z, Cheetham BMG & Homer JJ, "Voice Quality Assessment by Simulating GRBAS Scoring," in proceeding of UKSim-AMSS 11th European Modelling Symposium on Mathematical Modelling and Computer Simulation (EMS2017), Manchester, UK, November 2017.

[2] Xie Z., Gadepalli C. & Cheetham BMG, "Reformulation and Generalisation of the Cohen and Fleiss Kappas," *LIFE: International Journal of Health and Life-Sciences,* Vol 3 no 2, November 2017.

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

[4] Brennan RL & Prediger DJ, Coefficient Kappa: some uses misuses and alternatives, Educational and Psychological Measurement 41, 1981, pp.687-699.

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

[6] Hubert l., Kappa Revisited, Psychol Bull, 84, 1977, pp.289-297.

[7] Conger A.J., Integration and Generalisation of Kappas for Multiple Raters, Psychol Bull., 88, 1980, pp.322-328.

[8] Light R.J., Measures of response agreement for qualitative data: some generalisations and alternatives, Psychol Bull, 76, 1971, pp.365-377.

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

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

[11] Scott WA, Reliability of Content Analysis: the case of nominal scale coding, Public Opinion Quarterly XIX, 1955, pp. 321-325

[12] Aickin M., Maximum Likelihood Estimation of Agreement in the Constant Predictive Probability Model and its Relation to Cohen's Kappa, Biometrics, 46, 1990, pp. 293-302

[13] Lee Rodgers J. & Nicewander W.A., Thirteen ways to look at the correlation coefficient, The American Statistician. 42(1), 1998, pp. 59-66.

[14] Bland J.M. & Altman D., Statistical methods for assessing agreement between two methods of clinical measurement, The lancet, 327(8476), 1986, pp.307-310.

[15] Koch G.G., Intraclass correlation coefficient. *Encyclopedia of Statistical Sciences*, 1982.

[16] Rödel E., 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), 1971, pp.429-30.

TABLE I. DISTRIBUTION OF SUBJECT SCORES ILLUSTRATING THE GWET PARADOX

Rater	Scores for subjects 1-20																			
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1

TABLE II. DISTRIBUTION OF SCORES FOR SUBJECTS LIKELY TO BE EASY TO SCORE

Rater	Scores for subjects 1-20																			
1	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
2	1	2	2	2	1	2	1	2	1	2	1	2	1	2	1	1	1	2	1	2