Revised on October 05, 2015

On Krippendorff's Alpha Coefficient

Kilem L. Gwet, Ph.D.
Statistical Consultant
Advanced Analytics, LLC
PO Box 2696
Gaithersburg, MD 20886-2696
gwet@agreestat.com

ABSTRACT

Krippendorff's alpha coefficient is a statistical measure of the extent of agreement among coders, and is regularly used by researchers in the field of content analysis. This coefficient is known to involve complex calculations, making the evaluation and its sampling variation possible only through resampling methods such as the bootstrap. In this paper, we propose a simple procedure for calculating Krippendorff's alpha that involves simple calculations similar to those needed to obtain the Pi coefficient of Fleiss (1971). We will propose a close expression for computing its variance, as well as a new interpretation based on the notion of weighting. Additionally, we will present some alternative agreement coefficients, which address the classical problem of the paradoxes associated with Cohen's Kappa, and described by Cicchetti and Feinstein (1990) and Feinstein and Cicchetti (1990).

1. INTRODUCTION

The data that we collect during a research project generally reflect our understanding of the topic under investigation and our interpretation of the phenomena being observed. Our judgment however, will often not be in total agreement with what our peers are doing. Such discrepancies can be detrimental to the integrity of scientific inquiries, which require that the data associated with the same subjects be comparable regardless of their sources, if they are of the same type. This property guarantees the reproducibility of scientific data, and must be a key objective of any data quality control program.

The inter-rater reliability coefficient is a statistical measure that quantifies the extent of agreement among observers. Numerous statistics have been proposed in the literature for achieving this quantification goal. Many of these statistics take the form $(p_a - p_e)/(1 - p_e)$, where p_a is the percent agreement and p_e, the percent chance agreement. These coefficients are generally based on a common formulation of the percent agreement, and differ noticeably on their formulation of the percent chance agreement.

Scott (1955) proposed the π (read Pi) coefficient for quantifying the extent of agreement between 2 raters. Cohen (1960) later criticized Scott's proposal on the ground that its chance-agreement does not use each rater's specific classification propensities into categories, and proposed the κ (read Kappa) coefficient. The kappa coefficient, all its flaws notwithstanding remains widely used across many research fields. Cicchetti and Feinstein (1990), Feinstein and Cicchetti (1990) described several methodological problems associated with Kappa. Other agreement coefficients often used by researchers in various fields of research, include the Krippendorff α (read alpha) (see Krippendorff 1980, 2004a), Brennan-Prediger (1981) coefficient, Gwet's AC_1 (2008*a*), and more. Banerjee et al. (1999) provide a good literature review.

The first goal of this paper is to provide a simple approach for computing and interpreting the Krippendorff's alpha coefficient and its variance for researchers who rely on it to quantify the extent of agreement among raters. The new and simplified approach will be illustrated using the data set provided by Hayes and Krippendorff (2007). The second goal of this paper is to present an alternative agreement coefficient that addresses the paradox problem of Cicchetti and Feinstein (1990).

2. THE KRIPPENDORFF'S ALPHA

The procedure for computing Krippendorff's alpha is often described in terms of coincidence tables and difference functions (see Krippendorff, 2007). We will stay away from these two concepts, and replace them with the notions of agreement tables and weights, which are more commonly-used in the inter-rater reliability literature. We will formulate the alpha coefficient using the standard kappa-like form similar to the weighted Kappa coefficient introduced by Cohen (1968).

Consider a reliability experiment where r observers classify n objects into q categories. As is often the case in practice, some observers may not rate all subjects. This situation creates the possibility of having missing values in the dataset. Since missing values are common in practice, a good agreement coefficient is expected to handle them properly.

The general form of the alpha coefficient is defined as follows,

$$\alpha = \frac{p_a - p_e}{1 - p_e}, \tag{1}$$

where p_a is the weighted percent agreement and p_e the weighted percent chance agreement. The weighted percent agreement and weighted percent chance agreement will be formally defined later in this section. A weight is typically assigned to each pair of categories (k, l) and takes a value between 0 and 1 that decreases as the seriousness of the disagreement represented by the 2 categories increases. When 2 observers classify a subject into the exact same category k, they are considered to have achieved full agreement. The weight w_{kk} associated with the pair (k, k) is then assigned the maximum value of 1. If the 2 observers classify a subject into 2 distinct categories $(k \neq l)$, then the pair is assigned a weight smaller than 1 representing the fraction of full agreement (or partial agreement) this particular disagreement is credited of. See Cohen (1968) for a discussion of the notion of weighting in the context of inter-rater reliability. The different weights proposed for the Krippendorff's alpha coefficient are defined in section 3.

To fix ideas, let us consider the reliability data described in Table 1 that was initially considered by Krippendorff (2007). This table summarizes the outcome of an experiment where 4 observers A, B, C, and D classified 12 units into one of 5 possible categories labeled as 1, 2, 3, 4, and 5. It appears that none of the 4 observers rated all 12 units, which has produced a table with missing ratings.

TABLE 1
Ratings assigned by 4 raters A, B, C, and D to 12 Units

Unit	Observers			
	A	B	C	D
1	1	1		1
2	2	2	3	2
3	3	3	3	3
4	3	3	3	3
5	2	2	2	2
6	1	2	3	4
7	4	4	4	4
8	1	1	2	1
9	2	2	2	2
10		5	5	5
11			1	1
12			3	

The extent of agreement among the 4 observers could be quantified with the Krippendorff's alpha coefficient as described in the following steps:

STEP 1: Constructing the Agreement Table

This first step consists of creating a table that displays the distribution of observers by unit and category similar to Table 2. A typical entry of Table 2 is r_{ik}, and represents the number of observers who classified unit i into category k. It follows from the row associated with unit 8 that 3 raters classified unit 8 into category 1. That is $r_{81} = 3$.

The "Total" column shows the number of observers who rated each individual unit. For any particular unit i, that number is denoted by r_{i+}. For the purpose of calculating the Krippendorff's alpha coefficient all units rated by a single observer or not rated at all by any observer must be excluded from the analysis. Consequently, unit #12 of Table 2, will not be considered when calculating the α coefficient (note that $r_{12+} = 1$).

TABLE 2
Distribution of raters by unit and by category based on Table 1 data

Unit	Category					
	1	2	3	4	5	Total
1	3	0	0	0	0	3
2	0	3	1	0	0	4
3	0	0	4	0	0	4
4	0	0	4	0	0	4
5	0	4	0	0	0	4
6	1	1	1	1	0	4
7	0	0	0	4	0	4
8	3	1	0	0	0	4
9	0	4	0	0	0	4
10	0	0	0	0	3	3
11	2	0	0	0	0	2
12	0	0	1	0	0	1
Average	0.818	1.182	0.909	0.455	0.273	**3.636**
Classification Probability (π_k)	0.225	0.325	0.25	0.125	0.075	1

STEP 2: Calculating the Average Number of Observers per Unit, and the Classification Probabilities.

- The average number of observers per unit denoted by $\bar{r}$ (read r bar), is obtained by averaging all numbers in the ``Total" column of Table 2 (except the number 1 associated with unit #12). That is,

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i = \frac{3+4+4+4+4+4+4+4+4+4+3+2}{11} = \mathbf{3.636}.$$

That is, on average, about 3.636 observers rated each unit.

- For a given category k, the probability π_k that a randomly selected observer will classify any given unit into category k, is calculated as the ratio of the average number of observers who classified a unit into category k and rated 2 units or more, to the overall average number of raters per unit. That is,

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{\bar{r}}, \qquad (2)$$

where n is the number of units rated by 2 observers or more. The last row of Table 2 shows all 5 classification probabilities $\pi_1, \pi_2, \pi_3, \pi_4, and\ \pi_5$.

STEP 3: Calculating the overall percent agreement p_a

The overall percent agreement associated with Krippendorff's α coefficient is obtained as $p_a = (1 - 1/n\bar{r})p'_a + 1/n\bar{r}$, where p'_a is calculated by averaging all n unit-level percent agreement values $p_{a|i}$. These elements are defined as follows:

$$p'_a = \frac{1}{n}\sum_{i=1}^{n} p_{a|i},\ where\ \ p_{a|i} = \sum_{k=1}^{q}\frac{r_{ik}(\bar{r}_{ik+} - 1)}{\bar{r}(r_i - 1)},\ and\ \bar{r}_{ik+} = \sum_{l=1}^{q} w_{kl}r_{il}. \qquad (3)$$

The quantity $\bar{r}_{ik+}$ represents the weighted count of observers who classified unit i into any category that represents an agreement or a partial agreement with category k, as defined by the set of weights being used. This quantity can be calculated only after specifying the set of weights to be used. At this stage, we like to make the following 2 comments:

(i) When there are no missing ratings, the quantity p'_a represents the percent agreement that is associated with most kappa-like agreement coefficients known in the literature, including Fleiss' Kappa (Fleiss, 1971). The Krippendorff's alpha is based on a percent agreement p_a that always exceeds p'_a, although the difference decreases proportionally to the number of units in the sample, and the number of observers participating in the study.

(ii) Weighting the count of observers who classified unit i into category k and its affiliates (determined by the weights), amounts to counting all observers who classified unit i into category k, and a fraction of those who classified unit i into any other category with a non-zero weight with category k.

Krippendorff (2007) has proposed several "metric differences," each of which is associated with a specific set of weights as shown in section 3.

STEP 4: Calculating the overall percent chance agreement p_e

The percent chance agreement quantifies how often you would normally expect 2 randomly selected observers to agree if the scoring is performed randomly according to the observed classification probabilities. That is,

$$p_e = \sum_{k,l}^{n} w_{kl}\pi_k\pi_l. \quad (4)$$

The above expression is similar to the percent chance agreement that Gwet (2010) proposed for the weighted version Fleiss' generalized kappa, after adapting the methodology of Janson and Olsson (2001).

STEP 5: Computing the Krippendorff's alpha coefficient

Compute the α coefficient according to equation (1) using the percent agreement p_a and the percent chance agreement p_e given by equations (3) and (4).

After computing the Alpha coefficient, you may also compute its standard error using the variance expression shown in section A of the appendix. This expression is implemented in the AgreeStat 2015.4Excel program discussed in section 5.

3. THE WEIGHTS BASED ON KRIPPENDORFF'S METRIC DIFFERENCES

For any 2 categories k, and l the associated weight w_{kl} is defined as $w_{kl} = 1 - \delta_{kl}$, where δ_{kl} is the Krippendorff's metric difference (possibly standardized to obtain a number between 0 and 1). Several weights are available, and only the type of data (i.e. ratings) being analyzed determines the appropriate weight to be used.

- Identity Weights (I-weights)

  When dealing with nominal ratings, it is recommended to use the identity weights defined as follows:

  $$w_{kl} = \begin{cases} 1, & if\ k = l, \\ 0, & otherwise. \end{cases}$$

  Only the classification of units into the exact same category by 2 raters is considered as agreement, and there is no provision for partial agreement. The use of I-weights actually leads to the unweighted analysis.

- Ordinal Weights (O-weights)

  For ordinal ratings, Krippendorff (2007) has described a weighting scheme that depends on the number of subjects, and the distribution of raters by subject and category. Such weights do not have a straightforward interpretation since they are sample-dependent. They may also have an adverse impact on the variance of the alpha coefficient, in addition to making its standard error estimation more complex.

  I propose the following simpler weights in case of ordinal ratings:

$$w_{kl} = \begin{cases} 1, & if\ k = l \\ 1 - \dfrac{\#\{(i,j),\ \min(k,l) \le i < j \le \max(k,l)\}}{w_{max}}, & if\ k \neq l. \end{cases}$$

Note that, $\#\{(i,j),\ \min(k,l) \le i < j \le \max(k,l)\}$ represents the number of pairs (i,j) with $i < j$, which can be formed with numbers between $\min(k,l)$ and $\max(k,l)$, and w_{max} its maximum value over all values of k, and l. One may note that $\#\{(i,j), \min(k,l) \le i < j \le \max(k,l)\}$ represents the number of combinations of $\max(k,l) - \min(k,l) + 1$ objects taken 2 at a time.

- Quadratic Weights (Q-weights)

  When dealing with interval data, one would use the following quadratic weights that correspond to Krippendorff's interval metric differences:

$$w_{kl} = \begin{cases} 1, & if\ k = l, \\ 1 - \dfrac{(k-l)^2}{\max\limits_{a,b}(b-a)^2} & if\ k \neq l. \end{cases}$$

- Ratio Weights (R-weights)

  The ratio weights to be used with ratio data are defined as follows:

$$w_{kl} = 1 - [(k-l)/(k+l)]^2$$

- Circular Weights (C-weights)

  Let q_{max} and q_{min} be respectively the largest and smallest values of the scoring scale, and $U = q_{max} - q_{min} + 1$. The circular weights are defined as follows:

$$w_{kl} = \begin{cases} 1 - \left(sin\left[\dfrac{180(k-l)}{U}\right]\right)^2, & if\ the\ sine\ function's\ argument\ is\ in\ degrees, \\ 1 - \left(sin\left[\dfrac{\pi(k-l)}{U}\right]\right)^2, & if\ the\ sine\ function's\ argument\ is\ in\ radians. \end{cases}$$

- Bipolar Weights (B-weights)

  The Bipolar weights are defined for any 2 categories k, and l as follows:

$$w_{kl} = 1 - \frac{(k-l)^2}{(k+l-2q_{min})(2q_{max}-k-l)}\ .$$

4. GWET'S AC_1 COEFFICIENT

The Krippendorff's alpha coefficient has several advantages, which include the ability to handle missing data, and to be loaded with various sets of weights for different data types. However, this coefficient does not specifically address the classical paradox problem raised by Cicchetti and Feinstein (1990), as well as Feinstein and Cicchetti (1990). These authors have described the following two paradoxes with respect to the use of Kappa, which are in reality equally valid for all kappa-like agreement coefficients that use the same percent chance agreement of equation (4):

- With a large value associated with the percent chance agreement p_e, the correction procedure will convert a high percent agreement into a relatively low kappa value.
- In case of 2 raters, unbalanced marginal totals produce higher values of kappa than more balanced totals.

These problems were extensively discussed by Gwet (2008*a*), and the AC_1 coefficient proposed as a paradox-robust alternative to kappa. The weighted version of AC_1 (also referred to as AC_2), which is used for ordinal, interval, and ratio data, were discussed by Gwet (2010) in the case where there is no missing rating. Weighted versions of other coefficients, including Fleiss' generalized kappa (1971), Conger (1980), Brennan and Prediger (1981) based on the methods of Berry and Mielke (1988), Janson and Olsson (2001), and Janson and Olsson (2004), are also discussed by Gwet (2010). In this section, I will present a version of the weighted AC_1 that can handle missing values, while avoiding the paradoxical behavior of kappa, Fleiss' generalized kappa and the likes.

The general form of AC_2 is $\gamma_2 = (p_a - p_e)/(1 - p_e)$, where p_a is the weighted percent agreement, and p_e the weighted percent chance agreement, which are both defined differently from Krippendorff's. The percent agreement associated with AC_2 is obtained by averaging all[1] unit-level percent agreement values $p_{a|i}$, defined as,

$$p_{a|i} = \sum_{k}^{q} \frac{r_{ik}(\bar{r}_{ik+} - 1)}{r_i(r_i - 1)}. \qquad (5)$$

The form taken by this unit-level percent agreement is justified. It represents the probability that 2 observers, randomly selected among the r_i observers who rated unit i, agree about its classification. For a given category k, the first observer would be selected with probability r_{ik}/r_i, while the second should be selected with probability $(\bar{r}_{ik+} - 1)/(r_i - 1)$. Using $\bar{r}_{ik+}$ (see equation (3)) as opposed to r_{ik} is necessary to account for the partial agreement situations due to all these categories with a non-zero weight with category k (i.e. all category k affiliates).

The percent chance-agreement associated with AC_2 is calculated as follows:

[1] You would actually average over all units that were rated by 2 observers or more, which is also what is done with Krippendorff's alpha.

$$p_e = \frac{T_w}{q(q-1)} \sum_{k=1}^{q} \pi_k (1-\pi_k)\,, where\ T_w = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl},\ and\ \pi_k = \frac{1}{n}\sum_{i=1}^{n} \frac{r_{ik}}{r_i}. \qquad (6)$$

π_k represents the probability of classification into category k. The percent chance agreement associated with AC_2 is defined in such a way that it increases as the classification probabilities get closer to $1/q$ (a situation expected with random scoring). As the classification probabilities stray away from these uniform values, the weighted percent chance agreement decreases. Gwet (2008*a*) argues that the percent chance agreement should not use the observations as if they were all generated under the hypothesis of independence or by pure chance as done with kappa and other coefficients. At best, only a small portion of the data might have been generated by a mechanism susceptible to produce agreement by chance. The distribution of raters by category must suggest the extent to which chance agreement can be expected.

Note that the Krippendorff's α coefficient is solely based on units that are rated by 2 observers or more. However, Gwet's AC_1 and AC_2 coefficients consider units rated by 2 observers or more only for the purpose of calculating the percent agreement p_a; the percent chance agreement p_e is based on all units rated by one observer or more. Expressions for computing the standard error of the AC_2 coefficients are given in the appendix.

5. THE AGREESTAT PROGRAM

AgreeStat is an Excel VBA (Visual Basic for Applications) program that allow researchers to perform statistical analysis on the extent of agreement among multiple raters. It implements the Krippendorff's alpha with all of its weights, Gwet's AC_1 and AC_2 coefficients, Fleiss Kappa, and more, and can be downloaded at the following URL http://agreestat.com/agreestat.html. A unique feature of AgreeStat is the possibility for a researcher to specify custom weights to best reflect the special nature of certain disagreement situations.

The dataset shown in Table 1 was analyzed with AgreeStat and the results are shown in Table 3. The coefficient names appear in the first column, while the second column contains the coefficient estimates. Columns 3 and 4 contain the standard errors and 95% confidence intervals associated with the coefficients, and calculated with respect to the sampling of subjects only. That is, the observer sample is considered fixed and not subject to any sampling variability. This statistical inference is performed conditionally upon the observer sample, and allows a researcher to infer to the subject population only. For inference to both populations of subjects and observers simultaneously, one needs to use the standard errors and confidence intervals shown in columns 5 and 6.

Researchers dealing with inter-coder reliability have a tendency to consider inference with respect to the population of subjects only. This may well be all what is needed. After all, if you want to know the extent to which 2 coders agree, and you are interested exclusively in the coding of these 2 particular individuals, then the population of subjects is the only population of inference you should be concerned about. However, if the coders who participate in the reliability experiment are seen as a representative sample of a bigger population of possible coders, then you must consider not one, but 2 populations of inference, which are the population of subjects, and the population of coders. In this case, your statistical procedure must evaluate the precision of the agreement coefficients with respect to the 2 sources of variability due to the sampling of subjects and the sampling of coders. Statistical inference will no longer be conditional upon a fixed observer sample. Instead, it will be unconditional, with no fixed sample. Gwet (2008*b*) discusses this notion of unconditional standard error as it applies to inter-rater reliability coefficients.

AgreeStat evaluates the standard error with respect to both the sampling of subjects, and that of coders, by summing the subject variance and the coder variance, and taking the square root of the total variance. The subject variance is calculated using the expressions given in the appendix of this paper, while the coder variance component is calculated using the jackknife method as discussed in Gwet (2010).

TABLE 3
A Partial Output of AgreeStat Based on Table 1 Data

Method	Coefficient	Inference/Subjects		Inference/Subjects & Raters	
		Std. Err	95% C.I.	Std. Err	95% C.I.
Conger's Kappa	0.7628	0.1492	0.435 to 1	0.1898	0.345 to 1
Gwet's AC1	0.7754	0.1429	0.461 to 1	0.1814	0.376 to 1
Fleiss' Kappa	0.7612	0.1530	0.424 to 1	0.1945	0.333 to 1
Krippendorff's Alpha	0.7434	0.1455	0.423 to 1	0.1950	0.314 to 1
Brenann-Prediger	0.7727	0.1447	0.454 to 1	0.1838	0.368 to 1
Percent Agreement	0.8182	0.1256	0.542 to 1	0.1549	0.477 to 1

An agreement coefficient may have a good precision level with respect to the subject sample only, and a low precision level with respect to the sampling of both subjects and coders. This would be an indication that the number of coders recruited for the experiment was too small to produce an extent of agreement that can be projected to the entire universe of coders. The researcher will then have to either increase the number of coders in the experiment, or restrict statistical inference to the subject population, and the specific sample of coders at hand.

I now consider the dataset discussed by Hayes and Krippendorff (2007). This dataset contains the ratings that 5 observers assigned to 40 newspaper articles with respect to their coverage of a challenger running for political office against an incumbent. The observers had to rate the article with respect to what its tone suggested, and classify it in one of the following 4 categories:

- 0: The challenger is a sure loser
- 1: The challenger is somewhat competitive
- 2: The challenger is competitive
- 3: The challenger is a likely winner

None of the 5 observers rated all 40 articles, and some of them generated as many as 18 missing values. AgreeStat 2015.4 handles these missing values with no problem. Table 4 contains a partial output of AgreeStat 2015.4, where the ratings are treated as nominal data. That is, the Identity weights are used to assign a weight of 1 to the perfect agreement situations, and a weight of 0 to all situations of disagreements. It appears that all agreement coefficients yield low inter-rater reliability values, all being around 0.50.

TABLE 4
A Partial Output of AgreeStat 2015.4for the
Hayes-Krippendorff 2007 Dataset – Nominal Data

		Inference/Subjects		Inference/Subjects & Raters	
Method	Coefficient	Std. Err	95% C.I.	Std. Err	95% C.I.
Conger's Kappa	0.4726	0.0844	0.302 to 0.643	0.1252	0.219 to 0.726
Gwet's AC_1	0.5093	0.0654	0.377 to 0.642	0.1072	0.292 to 0.726
Fleiss' Kappa	0.4697	0.0696	0.329 to 0.61	0.1165	0.234 to 0.705
Krippendorff's Alpha	0.4765	0.0676	0.34 to 0.613	0.1191	0.236 to 0.717
Brenann-Prediger	0.5000	0.0661	0.366 to 0.634	0.1092	0.279 to 0.721
Percent Agreement	0.6250	0.0496	0.525 to 0.725	0.0819	0.459 to 0.791

Table 5 contains a partial output of AgreeStat 2015.4, where the ratings are treated as interval data. That is the Quadratic Weights, which correspond to Krippendorff's interval metric differences are used with all agreement coefficients. The coefficients resulting from the use of Q-weights are substantially higher than those based on I-weights. The chance-corrected measures range from 0.7499 for Fleiss' kappa to 0.8476 for Gwet's AC_1.

TABLE 5
A Partial Output of AgreeStat 2015.4for the
Hayes-Krippendorff 2007 Dataset – Interval Data

Method	Coefficient	Inference/Subjects	Inference/Subjects &

				Raters	
		Std. Err	95% C.I.	Std. Err	95% C.I.
Conger's Kappa	0.7536	0.0521	0.648 to 0.859	0.0610	0.63 to 0.877
Gwet's AC1	0.8476	0.0279	0.791 to 0.904	0.0347	0.777 to 0.918
Fleiss' Kappa	0.7499	0.0518	0.645 to 0.855	0.0616	0.625 to 0.874
Krippendorff's Alpha	0.7574	0.0424	0.672 to 0.843	0.0643	0.627 to 0.887
Brenann-Prediger	0.8250	0.0296	0.765 to 0.885	0.0374	0.749 to 0.901
Percent Agreement	0.9514	0.0082	0.935 to 0.968	0.0104	0.93 to 0.972

6. CONCLUDING REMARKS

In this paper, we have proposed a simple approach that we expect will make the computation of Krippendorff's alpha coefficient no more intimidating than the computation of any other agreement coefficient in the literature. We have also presented the alpha in a form that makes it more comparable to other known coefficients, in addition to proposing a close expression for calculating its variance. We derived this variance expression based on a linearized version of the alpha coefficient, which is valid for a reasonably large number of subjects.

We have also introduced the AC_1 (for nominal data) and AC_2 coefficients, which can also handle missing values, in addition to addressing the paradox problem associated with Cohen's kappa. AC_2 could use all weights based on the Krippendorff's metric differences to accommodate various data types.

We have briefly described the AgreeStat 2015.4program that can be used to compute various agreement coefficients, along with their standard errors and confidence intervals. It may also be used to conduct statistical inference conditionally upon the rater sample, or unconditionally. Unconditional inference has the advantage to project the analysis results to both universes of subjects and raters.

7. REFERENCES

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

Brennan, R. L., and Prediger, D. J. (1981). Coefficient Kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Cicchetti, D. V., and Feinstein, A. R. (1990). High Agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43,* 551-558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213-220.

Conger, A. J. (1980). Integration and Generalization of Kappas for Multiple Raters. *Psychological Bulletin}, 88,* 322-328.

Feinstein, A. R., and Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43,* 543-549.

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

Gwet, K. L. (2008*a*). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61,* 29-48.

Gwet, K. L. (2008*b*). Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika, 73,* 407-430.

Gwet, K. L. (2010). *Handbook of inter-rater reliability (2nd edition).* Maryland, US: Advanced Analytics, LLC

Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1,* 77-89.

Janson, H., and Olsson, U. (2001). A Measure of Agreement for Interval or Nominal Multivariate Observations. *Educational and Psychological Measurement, 61,* 277-289.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*, chapter 12. Sage, Beverly Hills, CA

Krippendorff, K. (2004a). *Content Analysis: An Introduction to its Methodology*, second edition, chapter 11. Sage, Thousand Oaks, CA

Krippendorff, K. (2007). Computing Krippendorff's Alpha-Reliability. http://www.asc.upenn.edu/usr/krippendorff/webreliability.doc

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly, XIX,* 321-325.

APENDIX

A. THE STANDARD ERROR OF KRIPPENDORFF'S ALPHA

The standard error of Krippendorff's alpha coefficient is calculated as the square root of its variance, whose general form is given by equation[2] (7).

$$v(\alpha) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\alpha_i^* - \alpha)^2, \qquad (7)$$

where α_i^* (the subject-level alpha value) is defined as follows:

$$\begin{cases} \alpha_i^* = \alpha_i - 2(1-\alpha)\,(p_{e|i} - p_e)/(1-p_e), \\ \alpha_i = (p_{a|i} - p_e)/(1-p_e)\,, \end{cases}$$

and the subject-level percent agreement $p_{a|i}$ and percent chance agreement $p_{e|i}$ are respectively given by,

$$p_{a|i} = \sum_{k}^{q}\frac{r_{ik}(\bar{r}_{ik+}-1)}{\bar{r}(r_i-1)} - p_a(r_i-\bar{r})/\bar{r}\,,$$

$$p_{e|i} = \sum_{k=1}^{q}\bar{\pi}_k\, r_{ik}/\bar{r} - p_e\,(r_i-\bar{r})/\bar{r},\ where\ \bar{\pi}_k = (\bar{\pi}_{k+} + \bar{\pi}_{+k})/2\,,$$

and $\bar{\pi}_{k+}$ and $\bar{\pi}_{+l}$ are weighted classification probabilities in categories k and l respectively, and defined as follows:

$$\bar{\pi}_{k+} = \sum_{l=1}^{q} w_{kl}\pi_l,\ \ and\ \ \bar{\pi}_{+l} = \sum_{k=1}^{q} w_{kl}\pi_k$$

Note that all known weight matrices are symmetrical, and for such weights $\bar{\pi}_{k+} = \bar{\pi}_{+k}$. However, AgreeStat 2015.4allows for the use of custom weights with no constraint of symmetry. Although asymmetrical weights would be difficult to interpret, they may technically be used, and would lead to different weighted classification probabilities $\bar{\pi}_{k+}$ and $\bar{\pi}_{+k}$. Also note that just like Krippendorff's coefficient itself, its variance is calculated based solely on subjects that are rated by 2 raters or more. All subjects rated by a single rater must be excluded from the variance calculation altogether.

[2] Calculating this variance of the alpha coefficient requires all subjects rated by a single rater to be excluded first. Only subjects rated by 2 raters or more must be used.

B. THE STANDARD ERROR OF GWET'S AC_2 COEFFICIENT

The standard error of the AC_2 coefficient is the square root of its variance given by equation (8).

$$v(\gamma_2) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\gamma_{2|i}^{*} - \gamma_2\right)^2, \tag{8}$$

where γ_2 is the AC_2 coefficient, and $\gamma_{2|i}^{*}$ the unit-level coefficient, which will now be defined. Let $\gamma_{2|i}$ be the AC_2 coefficient based on unit i only and the overall percent chance agreement p_e. The quantity $\gamma_{2|i}$ is defined as follows:

$$\gamma_{2|i} = \begin{cases} (n/n')\,(p_{a|i} - p_e)/(1-p_e), & if\ r_i \geq 2, \\ 0, & otherwise, \end{cases}$$

where n is the total number of units being rated, and $n'$the number of units rated by 2 observers or more, and $p_{a|i}$ is defined as follows:

$$p_{a|i} = \sum_{k}^{q}\frac{r_{ik}(\bar{r}_{ik+} - 1)}{r_i(r_i - 1)}.$$

Let $\gamma_{2|i}^{*}$ be the bias-adjusted AC_2 coefficient based on unit i, and defined as follows:

$$\gamma_{2|i}^{*} = \gamma_{2|i} - 2(1-\gamma_2)\frac{p_{e|i} - p_e}{1-p_e},$$

where the unit-level percent chance agreement $p_{e|i}$ is defined as,

$$p_{e|i} = \frac{T_w}{q(q-1)}\sum_{k=1}^{q}\pi_k\left(1 - \frac{r_{ik}}{r_i}\right), where\ \ T_w = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}.$$