

Agree or Disagree? A Demonstration of An Alternative Statistic to Cohen’s Kappa for Measuring the Extent and Reliability of Agreement between Observers

Qingshu Xie (MacroSys, LLC)

1. Introduction

Agreement analysis is an important tool that has been widely used in medical, social, biological, physical and behavioral sciences. Though there are many different ways of measuring agreement between observers or raters, they generally fall into two broad categories: the classical descriptive approach and the modeling approach. The classical descriptive approach includes all of the summary measures that range from measures of agreement between two observers to generalized measures of agreement among multiple observers. Summary measures are more often used than the modeling approach in practice because of their seeming simplicity. However, there are plenty examples of misinterpretation and misuse of summary agreement measures (e.g. Brennan & Prediger, 1981; Ludbrook, 2002; Maclure & Willett, 1987). Different measures of interrater reliability often lead to conflicting results in agreement analysis with the same data (e.g. Zwick, 1988).

Cohen’s (1960) kappa is the most used summary measure for evaluating interrater reliability. According to the Social Sciences Citation Index (SSCI), Cohen’s (1960) seminal paper on kappa is cited in over 3,300 articles between 1994 and 2009 (Zhao, 2011). Its citations are more than 10 times that of the second most popular agreement index Scott’s π (1955). A search of “Kappa AND Statistic” in Medline database turned out 2,179 citations during 1980 – 2010 (Kingman, 2011). Cohen’s kappa is widely introduced in textbooks and is readily available in various statistical software packages such as SAS, Stata and SPSS. Despite its popularity, Cohen’s kappa is not without problem.

This paper compares Cohen’s kappa (κ) and Gwet’s (2002a) AC_I (γ), with Bennett, Albert & Goldstein’s (1954) S as a reference for comparison. It aims to demonstrate that Gwet’s AC_I is a better alternative to Cohen’s kappa in agreement analysis with nominal data. This paper uses binary data of two raters as an example in the following analysis. Note that both Cohen’s kappa (κ) and Gwet’s AC_I (γ) are generalized for agreement analysis with multiple raters and multiple categories (Conger, 1980; Gwet, 2008).

2. What is Cohen’s Kappa?

An overall agreement rate is the ratio of the number of cases on which two raters are in agreement to the total number of cases in the analysis. Some agreement may happen by chance because of at least one rater’s guessing or randomly rating a case. It is desirable that an agreement indicator eliminates chance agreement so that interrater agreement is reliably measured.

Table 1 Distribution of N subjects by rater and response category

		Rater A			
		Category 1 (Yes)	Category 2 (No)	Total	
Rater B	Category 1 (Yes)	a	b	a+b	$P_{1.}=(a+b)/N$
	Category 2 (No)	c	d	c+d	$P_{2.}=(c+d)/N$
	Total	a+c	b+d	N	
		$P_{.1}=(a+c)/N$	$P_{.2}=(b+d)/N$		

Prior to discussion of kappa and AC_1 , several general definitions are provided since they are used in the rest of the paper. These include observed agreement, prevalence index (PI), bias index (BI), and the general definition of a chance-corrected agreement measure.

Observed agreement is defined as

$$P_a = \frac{a + d}{N} \quad (1)$$

According to Byrt, Bishop and Carlin (1993), PI measures the differences in the probabilities of “Yes” and “No” responses, i.e. the difference between $(P_{1.} + P_{.1})/2$ and $(P_{2.} + P_{.2})/2$, which are the best estimates for prevalence in the population; BI measures the difference in probabilities of “Yes” responses between two raters, i.e. $(P_{1.} - P_{.1})$. By transformation, they are defined as below:

$$PI = \frac{a - d}{N}, \quad \text{and} \quad (2)$$

$$BI = \frac{b - c}{N}. \quad (3)$$

All chance corrected agreement measures can be defined in the following general form:

$$\text{Agreement Coefficient} = \frac{P_a - P_e}{1 - P_e}. \quad (4)$$

P_e is expected agreement by chance or chance agreement. Different agreement measures differ in their definitions of chance agreement.

Cohen’s Kappa is designed to correct for chance agreement (1960). It is defined as

$$\kappa = \frac{P_a - P_{e(k)}}{1 - P_{e(k)}}. \quad (5)$$

In Cohen’s kappa, chance agreement is defined as the sum of the products of marginal distributions, i.e.

$$P_{e(k)} = P_{.1} * P_{1.} + P_{.2} * P_{2.}. \quad (6)$$

Table 2 Popular reference levels of strength of agreement measured by Cohen’s kappa

Landis and Koch (1977) Biometrics article		Altman, DG (1991) Textbook		Fleiss et al (2003) Textbook	
Cohen's Kappa	Strength of Agreement	Cohen's Kappa	Strength of Agreement	Cohen's Kappa	Strength of Agreement
0.81 – 1.00	excellent	0.81 – 1.00	very good	0.75 – 1.00	very good
0.61 – 0.80	substantial	0.61 – 0.80	good	0.41 – 0.75	fair to good
0.41 – 0.60	moderate	0.41 – 0.60	moderate	< 0.40	poor
0.21 – 0.40	fair	0.21 – 0.40	fair		
0.00 – 0.20	slight	< 0.20	poor		
< 0.00	poor				

In the literature, there are several widely used reference levels of strength of agreement measured by Cohen's Kappa (see Table 2). Ludbrook (2002, p. 533) indicates that Landis & Koch’s (1977) approach “has no sound theoretical basis and can be positively misleading to investigators.” Unfortunately, while acknowledging that it is arbitrary, Oleckno (2008, pp. 475) still presents Landis and Koch’s suggested levels of kappa in a recent textbook for graduate students as “helpful” guidelines in “providing a general sense of the level of agreement beyond random

expectations.” Obviously, the three criteria in Table 2 do not agree with each other. Furthermore, as shown in the next section, these suggested levels of strength do not provide appropriate guidance on how to use kappa in agreement analysis since kappa is not a stable agreement measure and it varies with prevalence and bias dramatically at the same agreement rate.

3. Cohen’s Kappa is not a reliable general measure for interrater reliability but highly controversial

It is quite puzzling why Cohen’s kappa has been so popular in spite of so much controversy with it. Researchers started to raise issues with Cohen’s kappa more than three decades ago (Kraemer, 1979; Brennan & Prediger, 1981; Maclure & Willett, 1987; Zwick, 1988; Feinstein & Cicchetti, 1990; Cicchetti & Feinstein, 1990; Byrt, Bishop & Carlin, 1993). In a series of two papers, Feinstein & Cicchetti (1990) and Cicchetti & Feinstein (1990) made the following two paradoxes with Cohen’s kappa well-known: (1) A low kappa can occur at a high agreement; and (2) Unbalanced marginal distributions produce higher values of kappa than more balanced marginal distributions. While the two paradoxes are not mentioned in older textbooks (e.g. Agresti, 2002), they are fully introduced as the limitations of kappa in a recent graduate textbook (Oleckno, 2008). On top of the two well-known paradoxes aforementioned, Zhao (2011) describes twelve additional paradoxes with kappa and suggests that Cohen’s kappa is not a general measure for interrater reliability at all but a measure of reliability under special conditions that are rarely held.

Krippendorff (2004) suggests that Cohen’s Kappa is not qualified as a reliability measure in reliability analysis since its definition of chance agreement is derived from association measures because of its assumption of raters’ independence. He argues that in reliability analysis raters should be interchangeable rather than independent and that the definition of chance agreement should be derived from estimated proportions as approximations of the true proportions in the population of reliability data. Krippendorff (2004) shows that agreement measures can be defined as below:

$$Agreement = 1 - \frac{Observed\ Disagreement}{Expected\ Disagreement} = 1 - \frac{D_o}{D_e} \quad (7)$$

He mathematically demonstrates that kappa’s expected disagreement is not a function of estimated proportions from sample data but a function of two raters’ individual preferences for the two categories.

Still, some researchers strongly defend the use of kappa in agreement analysis. Vach (2005) argues that the dependence of Cohen’s kappa on the prevalence is not an issue at all. He distinguishes kappa’s dependence on prevalence into two types: “a dependence on the observed marginal prevalence and a dependence on the prevalence of a latent binary variable, representing the true status.” He argues that kappa’s dependence on observed prevalence is the purpose of kappa since it helps to improve the interpretation of agreement rates and that kappa’s dependence on the latent prevalence is negligible. It is true that appropriate correction for chance agreement is desirable. But the dramatically shifting of kappa with observed prevalence actually makes it difficult to interpret kappa. Oleckno (2008, pp. 479-480) indicates that kappa’s strong dependence on prevalence not only makes it difficult to interpret kappa but also adversely affect comparisons of kappa statistics across studies with different levels of prevalence of disorder.

When one rater’s ratings completely fall into one category, kappa is always equal to zero regardless of the level of agreement between two raters. This is because Cohen’s chance agreement is always equal to agreement rate in such situation. When two raters agree 100 percent in one category, Cohen’s kappa even becomes undefined. However, Kraemer, Periyakoil, and Noda (2004, p. 90) argue that “k=0 indicates either that the heterogeneity of the patients in the population is not well detected by the raters or ratings, or that the patients in the population are homogeneous” and that it is “not a flaw in kappa or any other measure of reliability, or a paradox.” But this argument is hardly convincing since we would expect some level of reliability when both professionally qualified raters agree at 70, 80 or 90 percent in their ratings in such situation.

In recognition of the limitations or problems with kappa, researchers have proposed many alternative agreement measures such as Krippendorff’s (1970) alpha (α) and Gwet’s (2002a) AC_1 (γ). Sometimes a newly invented measure may not be much more innovative. Krippendorff (2004) points out that at least five measures proposed as

better alternatives to Cohen's Kappa are actually equivalent of Bennett, Albert and Goldstein's S (1954). These include Holley & Guilford's G (1964), Maxwell's $R.E.$ (1970), Jason & Vegelius' C (1979), Brennan & Prediger's Kn (1981), and Perreault & Leigh's Ir (1989). In addition, as acknowledged by the authors themselves, Byrt, Bishop & Carlin's $PABAK$ (prevalence-adjusted bias-adjusted kappa) (1993) is the same as Bennett, Albert and Goldstein's S for binary data. The original formula for S is as below:

$$S = \frac{K}{K-1} \left(P_a - \frac{1}{K} \right). \quad (8)$$

It can be transformed into the general form of agreement measure:

$$S = \frac{P_a - \frac{1}{K}}{1 - \frac{1}{K}} \quad (9)$$

where expected agreement is $1/K$. When the data is binary, $S = 2P_a - 1$. In the next sections, Cohen's kappa, Gwet's AC_1 and Bennett, Albert and Goldstein's S are compared to show their different behaviors with prevalence and bias.

4. Two Paradoxes of Cohen's Kappa Explained – the effects of prevalence and bias

To overcome the effects of prevalence and bias on kappa agreement statistic, Byrt, Bishop and Carlin (1993) proposes $PABAK$ as an alternative measure. They find the relationship between Cohen's Kappa, their $PABAK$, and prevalence and bias in the formula below:

$$K = \frac{PABAK - PI^2 + BI^2}{1 - PI^2 + BI^2}. \quad (10)$$

Since $PABAK = 2P_a - 1$, equation (10) can be rewritten as

$$K = \frac{(2P_a - 1) - PI^2 + BI^2}{1 - PI^2 + BI^2}. \quad (11)$$

Equation (11) shows the following characteristics of kappa:

- (1) Holding PI and BI constant, Cohen's Kappa is an increasing function of agreement rate (P_a). That is, when marginal distributions are fixed, the higher the observed agreement rate, the greater the kappa.
- (2) When P_a is not equal 100%, holding prevalence constant, Cohen's Kappa is a nonlinear increasing function of the absolute value of bias index, which shows the effect of bias leading to the second paradox; holding bias constant, it is a nonlinear decreasing function of the absolute value of prevalence index, which shows the effect of prevalence leading to the paradox of high agreement but low kappa.
- (3) The combined effect of prevalence and bias on Cohen's Kappa depends on the relative magnitude of the two indexes.

The two paradoxes are due to kappa's dependence on marginal distributions (von Eye and von Eye, 2008). They are fundamentally due to the definition of chance agreement for kappa. By transformation, the general form of agreement statistics can be rewritten as:

$$\text{Agreement Coefficient} = 1 - \frac{1 - P_a}{1 - P_e}. \quad (12)$$

It is equivalent of equation (7). From equation (12), it is clear that, holding observed agreement rate (P_a) constant, agreement coefficient is a decreasing function of expected agreement by chance (P_e). If P_e moves erratically with the

changes of marginal distributions, the corresponding agreement coefficient is deemed to behave oddly. Therefore, how the chance agreement of an agreement measure is defined eventually determines its performance and differentiates it from other agreement measures.

5. A comparison of Cohen’s kappa, Gwet’s AC1 and Bennett, Albert and Goldstein’s S

Table 3 is a hypothetical distribution of observed agreement rates between two raters at different marginal distributions. These hypothetical agreement rates are used in producing the agreement statistics presented in the figures below. Since marginal distributions are known (in bold), chance agreement for kappa and AC_1 can be calculated.

Table 3 An Example of Observed Agreement Rates (2X2)

		Percent of Rater A's Rating of Category 1											
		0	10	20	30	40	50	60	70	80	90	100	
Percent of Rater B's Rating of Category 1	100	0	10	20	30	40	50	60	70	80	90	100	
	90	10	20	30	40	50	60	70	80	90	100	90	
	80	20	30	40	50	60	70	80	90	100	90	80	
	70	30	40	50	60	70	80	90	100	90	80	70	
	60	40	50	60	70	80	90	100	90	80	70	60	
	50	50	60	70	80	90	100	90	80	70	60	50	
	40	60	70	80	90	100	90	80	70	60	50	40	
	30	70	80	90	100	90	80	70	60	50	40	30	
	20	80	90	100	90	80	70	60	50	40	30	20	
	10	90	100	90	80	70	60	50	40	30	20	10	
	0	100	90	80	70	60	50	40	30	20	10	0	

5.1. Cohen’s kappa

As shown in equation (6), chance agreement in Cohen’s kappa is defined as the sum of the products of marginal distributions. It is naturally dependent on marginal distributions and leads to kappa’s dependency on marginal distributions. Figure 1 shows the erratic behavior of chance agreement in Cohen’s kappa. As shown in Figures 2 and 3, kappa chance agreement is near 0.5 when marginal distributions are close to 0.5. When there is marginal homogeneity, Figure 2 shows that the further away marginal distributions are from 0.5, which also means the increase of the absolute value of prevalence index (PI), the greater the chance agreement. When there is no marginal homogeneity, Figure 3 shows that the further away marginal distributions are from 0.5, which means the increase of the absolute value of bias index (BI), the smaller the chance agreement.

Figure 1 Distribution of chance agreement for Cohen's kappa by raters' marginal distributions

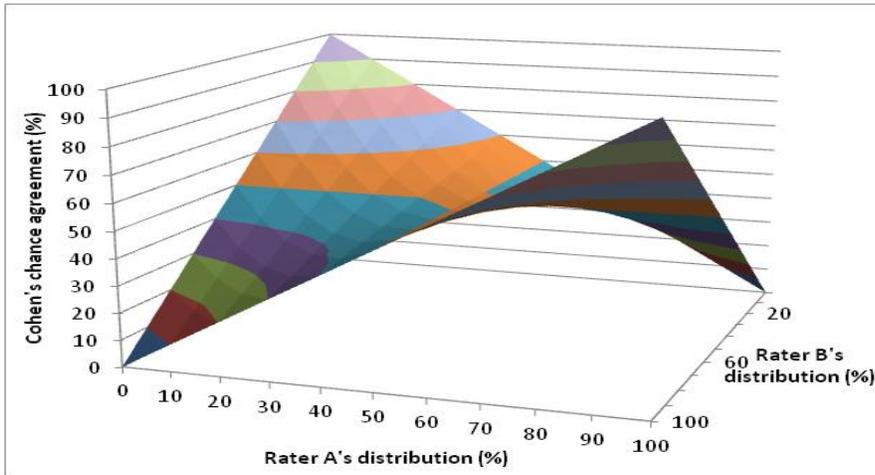


Figure 2 Chance agreement for Cohen's kappa with marginal homogeneity

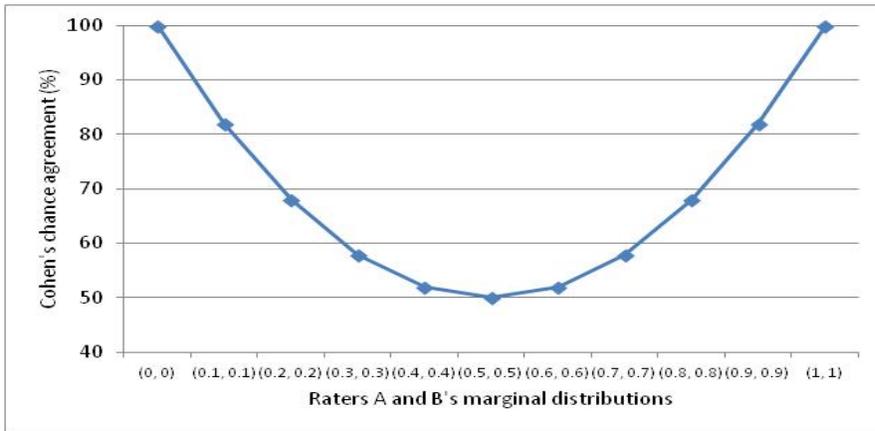
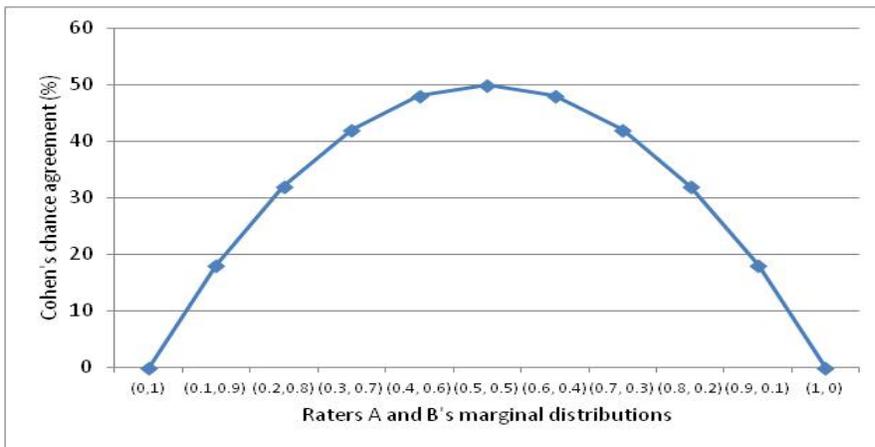


Figure 3 Chance agreement for Cohen's kappa without marginal homogeneity



As shown in equation (12), at a given observed agreement rate, agreement statistic is a decreasing function of chance agreement. The increase of chance agreement leads to the decrease of kappa; and the decrease of chance

agreement results in the increase of kappa. Figure 4 shows the effect of prevalence on the reduction of Cohen's kappa at fixed 90% observed agreement rate and 10% bias index. Cohen's kappa decreases with the increase of the absolute value of PI.

In Figure 5 each curve shows that at a fixed 40% agreement rate Cohen's kappa increases with the increase of the absolute value of BI. A comparison of the two curves indicates that at the same level of observed agreement rate and bias higher prevalence (i.e. the red curve) leads to lower kappa. Figure 6 shows that, with fixed prevalence, a higher observed agreement rate results in a higher kappa at different level bias.

Figure 4 Effect of prevalence on Cohen's kappa

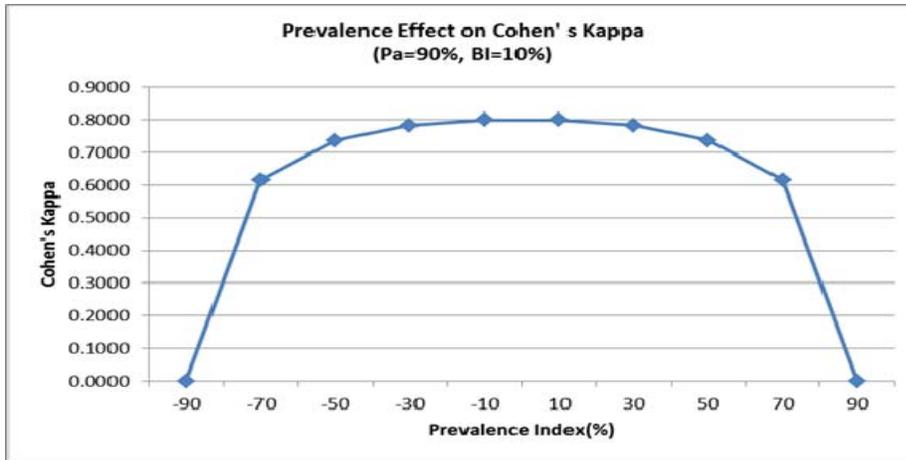


Figure 5 Effect of bias on Cohen's kappa

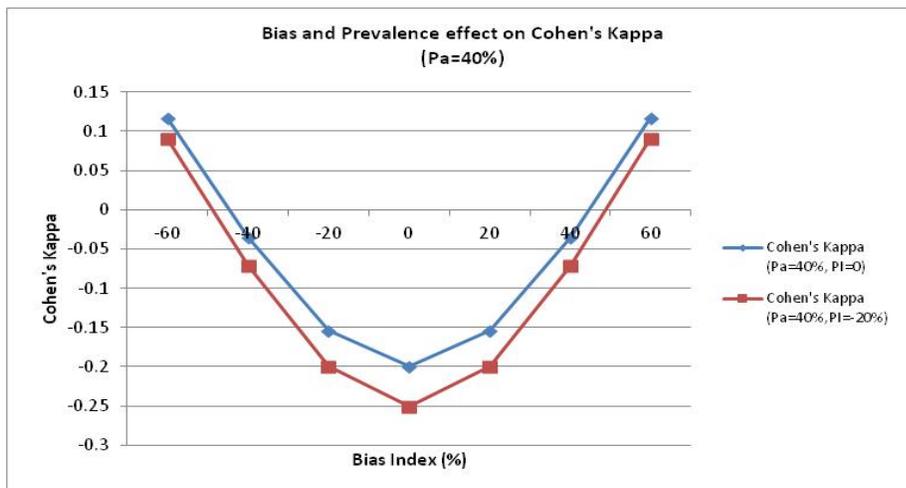
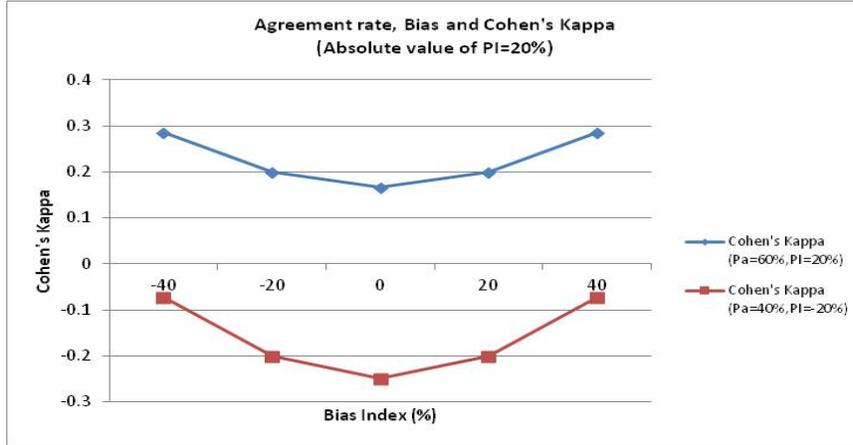


Figure 6 Relationship between agreement rate, bias and Cohen's kappa with fixed prevalence



5.2. Gwet's AC1 (γ)

Gwet's AC_1 is introduced as an alternative chance-corrected agreement measure to all existing agreement measures. Gwet (2008, p. 37) mathematically proves that AC_1 has "a smaller bias with respect to the 'true' agreement coefficient than all its competitors." Its definition of chance agreement is based on the premises that chance agreement does occur when at least one rater guesses or rates randomly and that ratings are not completely random but only an unknown proportion of ratings is random. Based on simulations, Gwet (2002b, p. 3) indicates that "a reasonable value for chance-agreement probability should not exceed 0.5." He further finds that chance agreement probability for Scott's (1955) π is from 0.5 to 1 and that chance agreement probability for Cohen's (1960) kappa can be any value between 0 and 1. This erratic behavior of chance agreement in kappa is already shown previously in Figure 1.

Gwet's AC_1 (γ) is defined as

$$AC_1 = \frac{P_a - P_{e(\gamma)}}{1 - P_{e(\gamma)}} \quad (13)$$

Where chance agreement $P_{e(\gamma)}$ is given by

$$P_{e(\gamma)} = 2\left(\frac{P_{1.} + P_{.1}}{2}\right)\left(1 - \left(\frac{P_{1.} + P_{.1}}{2}\right)\right) \quad (14)$$

Figure 7 shows the distribution of chance agreement for Gwet AC_1 by both raters' marginal distributions. Unlike chance agreement for both Scott's π and Cohen's kappa, chance agreement for Gwet's AC_1 is capped within 0 – 0.5. It is this limit that prevents erratic behavior of agreement statistic from happening in Gwet's AC_1 . In addition, Figure 8 shows that the pattern of chance agreement with marginal homogeneity for Gwet's AC_1 is reversed in comparison with that for Cohen's kappa as shown in Figure 2. Without marginal homogeneity, the curve for chance agreement for Gwet's AC_1 becomes a straight line (not shown here), suggesting that AC_1 is free from effect of bias.

Figure 7 Distribution of chance agreement for Gwet's AC_I by raters' marginal distributions.

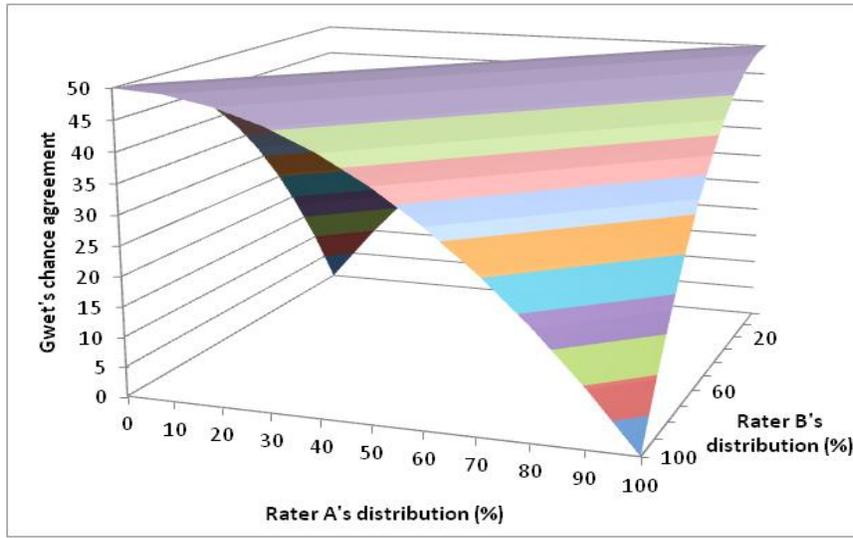


Figure 8 Chance agreement for Gwet's AC_I with marginal homogeneity

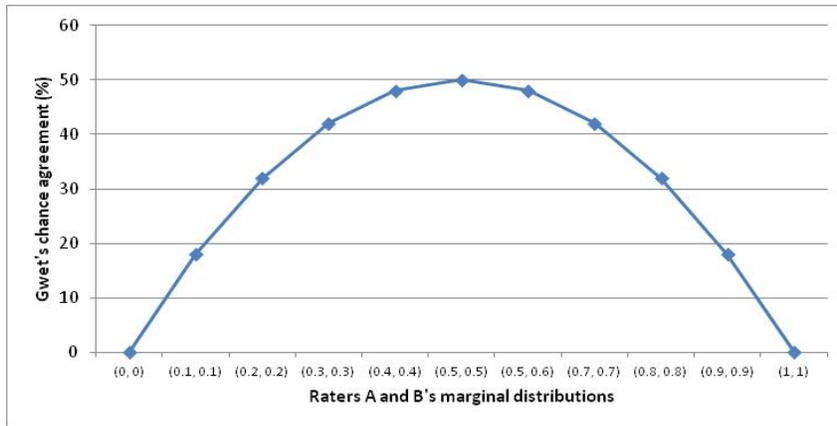


Figure 9 Gwet's AC_I with marginal homogeneity

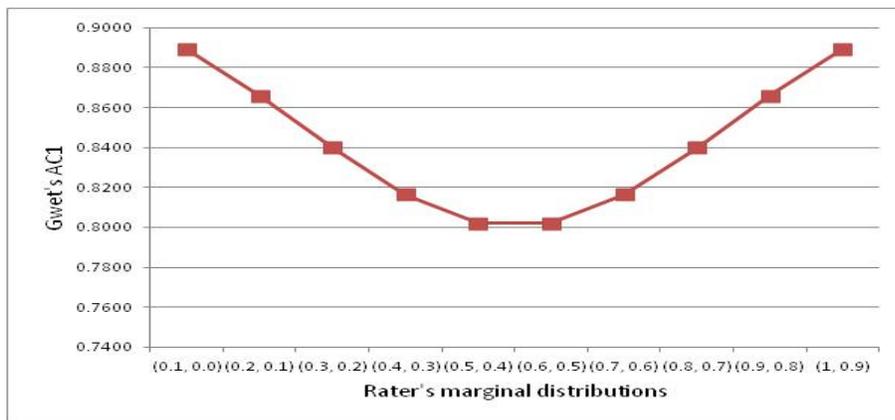


Figure 9 shows that Gwet's AC_I is still dependent on marginal distribution, but to a less extent, and in a reverse direction. It overcomes the paradox of high agreement low agreement statistic in Cohen's kappa. Based on the premises for AC_I , the dependence on marginal distributions is logical since we would expect a higher reliability when two qualified raters reach high agreement in observed ratings.

Using the definition of marginal distributions in Table 1, after equations (2) and (3) are introduced into equation (14), chance agreement for Gwet's AC_I can be expressed as

$$Pe(\gamma) = \frac{1}{2}(1 - PI^2) . \quad (15)$$

Thus, AC_I is transformed as below

$$AC_I = \frac{2P_a - 1 + PI^2}{1 + PI^2} . \quad (16)$$

Equation (16) suggests that Gwet's AC_I is related to prevalence but is not related to bias at all. This is because bias is already washed out in the definition of chance agreement for AC_I by using average marginal distributions as its estimates of proportions. While Cohen's kappa punishes raters who produce similar ratings or marginal distributions, such punishment does not exist in AC_I at all. On the contrary, raters with homogeneous marginal distributions and greater absolute value of PI are rewarded as shown in Figure 9. Gwet's AC_I overcomes the two well-known paradoxes with Cohen's kappa. The slight dependence on prevalence of AC_I is a desirable attribute of an agreement measure since it is in line with the common sense that we expect higher level of reliability when two professionally qualified raters reach similar or higher level of observed agreement.

5.3. Bennett, Albert & Goldstein's S

As shown in equation (9), Bennett, Albert & Goldstein's S is only a function of observed agreement rate and the number of categories for ratings or responses. It is not dependent on marginal distributions and is often referred to as a marginal free agreement index. S tends to underestimate interrater reliability because it assumes uniform marginal distributions and does not use observed proportions to estimate the true proportions in the population. This is especially obvious at the presence of prevalence. Since its correction for chance has nothing to do with the proportions in the population, as pointed out by Krippendorff (2004, p. 5), "it cannot indicate the reliability in the population of data."

Figure 10 Comparison of Gwet's AC_I , Cohen's kappa and Bennett et al's S

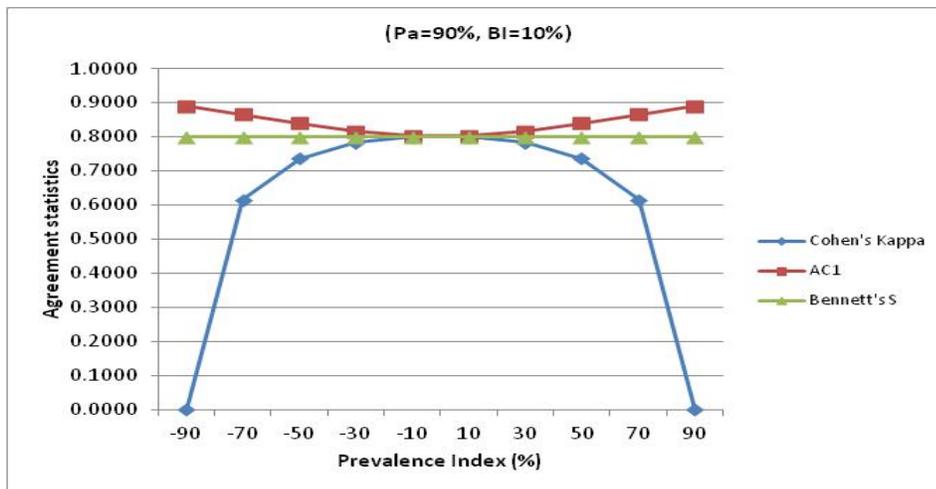


Figure 10 provides a comparison of Gwet's AC_1 , Cohen's kappa and Bennet et al's S. In the neighborhood of 0.5 marginal distributions or near zero prevalence, the three indicators provide similar magnitude of agreement statistics. However, they differ substantially with the increase of the absolute value of prevalence. It shows that AC_1 is a better alternative agreement measure to Cohen's kappa.

6. Conclusion

Cohen's kappa is a controversial agreement measure. It should be used with caution. If it is ever used, it should be reported with other indexes such as percent of positive and negative ratings, prevalence index, bias index, and test of marginal homogeneity. Considering its erratic behavior with extreme marginal distributions, it is a good practice to refrain from using it in such situation. Due to its lack of theoretical basis and controversy about kappa, it is not reliable to use the suggested levels of agreement strength on Cohen's Kappa to assess the reliability of an agreement study. Despite its popularity, Cohen's kappa should not be taken as a default indicator for interrater reliability. On the contrary, researchers need to look for other better alternatives.

Gwet's AC_1 is based on reasonable premises and is designed to overcome the limitations of existing agreement measures. The exercise in this paper demonstrates that it overcomes the well-known paradoxes in Cohen's kappa and it is not affected by bias at all. It is a better alternative to Cohen's Kappa. Since it is a relatively new measure, it is not yet readily available in standard software packages. But it can be programmed using standard statistical packages or can even be computed in Excel. It may always be a good idea to compute multiple measures in an agreement study.

References

- Agresti, A. (2002). *Categorical data analysis (2nd ed.)*. John Wiley & Sons, Inc., Hoboken, New Jersey, USA.
- Altman, D. G. (1991). *Practical Statistics for Medical Research*, Chapman & Hall, London.
- Bennett EM, Albert R, Goldstein AC (1954) Communications through limited response questioning. *Public Opinion Quarterly*, 18 (3), 303-308.
- Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41 (3), 687-699.
- Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology*, 46 (5), 423-429.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20 (1), 37-46.
- Cicchetti, D. V. & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43 (6), 551-558.
- Conger, A. J. (1980). Integration and generalization of kappa for multiple raters. *Psychological Bulletin*, 88 (2), 322-328.
- Feinstein, A. R. & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of the two paradoxes. *Journal of Clinical Epidemiology*, 43 (6), 543- 549.
- Fleiss J, Levin B., & Paik M (2003). *Statistical Methods for Rates & Proportions*, 3rd Ed. Wiley & Sons, New York.
- Gwet, K. L. (2002a). *Handbook of Interrater Reliability*. STAXIS Publishing Company.
- Gwet, K. L. (2002b). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Series : Statistical Methods for Inter-Rater Reliability Assessment, No. 1*. STAXIS Consulting.
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematics and Statistical Psychology*, 61, 29-48.
- Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement*, 24, 749-754.
- Janson, S., & Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research*, 14, 255-269.
- Kingman, A. (2011). Assessing Examiner Agreement: Why Reporting Kappa is NOT Enough. Presentation at The European Association of Dental Public Health (EADPH) 16th annual meeting, Rome, Italy, September 22-24, 2011.

- Krippendorff, K.. (1970). Bivariate agreement coefficients for reliability data. In E. R. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological Methodology 1970* (pp. 139-150). San Francisco, CA: Jossey Bass.
- Krippendorff, K.. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. Annenberg School for Communication Departmental Papers (ASC).
http://repository.upenn.edu/cgi/viewcontent.cgi?article=1250&context=asc_papers .
- Kraemer, H. C. (1979). Ramifications of a population model for k as a coefficient of reliability. *Psychometrika*, 44 (4), 461–472.
- Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2004). Kappa coefficients in medical research. In R. B. D'Agostino (Ed.): *Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies*. John Wiley & Sons, Ltd.
- Landis J. R., & Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33 (1), 159–174.
- Ludbrook, J. (2002). Statistical techniques for comparing measures and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology*, 29 (7), 527-536.
- Maclure, M., & Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126 (2), 161-169.
- Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry*, 130, 79–83.
- Oleckno, W. (2008). *Epidemiology: Concepts and Methods*. Long Grove, IL: Waveland Press, Inc., pp. 649.
- Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, 26, 135-148.
- Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, XIX, 321-325.
- Vach, W. (2005) The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, 58 (7), 655-661.
- Von Eye A., & von Eye M. (2008). On the Marginal Dependency of Cohen's κ . *European Psychologist*, 13 (4), 305–315.
- Zhao, X. (2011). When to use Cohen's K, If ever?, International Communication Association 2011 Conference, Boston, Massachusetts, U.S.A.
- Zwrick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, 103 (3), 374-378.