Impact of Rater Disagreement on Chance-Corrected Inter-Rater Agreement Indices with Equal and Unequal Marginal Proportions

David A. Walker
Northern Illinois University

This study examined the effect that equal free row and column marginal proportions, unequal free row and column marginal proportions, and the magnitude of rater disagreement had on eight agreement indices. In condition 1, when there were equal free row and column marginal proportions with no rater disagreement present, seven of the eight indices of agreement yielded very comparable results. In condition 2, when there were unequal free row and column marginal proportions and rater disagreement was ≤ .10, five of the eight indices of agreement tended to produce similar results. In conditions 3 and 4, when the marginals were not homogeneous and the amount of rater disagreement was > .10, there were three instances each of over-estimation and under-estimation. Thus, as cells B and C became less homogeneous, all of the inter-rater agreement indices studied, except for Cohen and Dice, were influenced via under- or over-estimation once rater disagreement was > .10. If rater disagreement was ≤ .10, 5 out of the 8 indices studied were not influenced by some degree of marginal heterogeneity.

In social science research, inter-rater agreement indices of categorical data for two raters have been studied extensively, and their strengths and weaknesses in various methodological situations reviewed in contexts such as classroom observations, political polling, psychological analysis, and content analysis (Bennett, Alpert, & Goldstein, 1954; Krippendorff, 2004; Riffe & Freitag, 1997; Zwick, 1988). Inter-rater agreement is conducted to verify that rater agreement exceeds, or does not, chance levels of agreement. The range of rater agreement is from -1.00 to +1.00, with +1.00 as total agreement, 0 as not better than chance that the raters would agree, and negative results indicate agreement worse than expected by chance due to random or systematic differences between raters such as rater bias or coding errors (Kassarjian, 1977; Linn & Gronlund, 2000; Sim & Wright, 2005).

In the literature pertaining to inter-rater agreement, various indices used with two raters and binary data emerge. All of these indices use a 2 x 2 agreement matrix, where the main diagonal (i.e., cells A and D) indicates the agreement level between the raters as either 00 or 11 and the off diagonal (i.e., cells B and C) indicates the level of disagreement between the raters as either 10 or 01.

		Rater 2		
		0	**1**	**Total**
Rater 1	**0**	Cell A	Cell B	$\mathbf{p_1}$
	1	Cell C	Cell D	$\mathbf{q_1}$
	Total	$\mathbf{p_2}$	$\mathbf{q_2}$	**n**

Figure 1. A 2 x 2 Matrix Configuration.

There are numerous indices for inter-rater agreement corrected for chance that can be applied to 2 x 2 tables with categorical data. However, in the scholarly literature (Fleiss, 1975; Hertzberg, Xu, & Haber, 2006; Krippendorff, 2004; Rae, 1988; Sirotnik, 1981; Übersax, 1987; Zwick, 1988), the following eight measures of agreement have been noted as common indices used and can be defined as "… proposed for categorical response data where such response is the assignment of the subject to one of κ mutually exclusive and exhaustive categories. [and] as a measure of agreement between multiple observations of a single subject" (Kraemer, 1979, p. 461).

The first chance corrected index for inter-rater agreement using a 2 x 2 table was proposed by Bennett et al. (1954) as Bennett's *S* coefficient, which requires the assumption of uniform marginals, where:

$$S = \frac{k}{k-1}(P_o - \frac{1}{k}) \tag{1}$$

where, P_o = observed agreement, where P_o = A + D; A = count from cell A, D = count from cell D; and k = number of response categories. Scott (1955) proposed Scott's pi coefficient or π, which requires the assumption of homogeneous marginals for the raters, where:

$$\pi = \frac{P_o - P_e}{1 - P_e} \ . \tag{2}$$

Note: Fleiss' intraclass correlation coefficient (1975) in a 2 x 2 situation is the same formula as Scott's π, with the assumption of equally distributed marginals where,

$P_{\mathrm{e}} = \sum_{i=1}^{k} \frac{(n_{i.} + n_{.i})^2}{2}$, is expected percentage of agreement based on chance, k = number of response categories. $n_{\mathrm{i.}}$ = observed row marginals for response i for rater 1, and $n_{\mathrm{.i}}$ = observed column marginals for response i for rater 2.

Cohen (1960) proposed Cohen's kappa coefficient or κ (1960), but did not have an assumption related to equally-distributed marginals, yet did assume that "… N objects categorized are independent; the assigners operate independently; and the categories are independent, mutually exclusive, and exhaustive" (Brennan & Prediger, 1981, p. 688).

$$\kappa = \frac{P_{\mathrm{o}} - P_e}{1 - P_e} \qquad (3)$$

where, A = count from cell A, B = count from cell B, C = count from cell c, D = count from cell D,
P_{o} = (A + D)/N is the observed agreement, N = number of observations, and
P_{e} = [(A + B)*(A + C)+(C + D)*(B + D)]/N^2 is the expected percentage of agreement based on chance.

Armitage, Blendis, and Smyllie (1966) proposed the standard deviation index, which is very similar to κ with no distributional assumption, but with the same assumptions of independence. Equations 5 to 7 do not have a distributional assumption, but also have the same independence assumptions as κ.

$$SD = \frac{(\mathrm{AD} - \mathrm{BC})(\mathrm{p}_1\mathrm{q}_1 + \mathrm{p}_2\mathrm{q}_2)}{2\mathrm{p}_1\mathrm{q}_1\mathrm{p}_2\mathrm{q}_2}\ ; \qquad (4)$$

where p_1 = (A+B), p_2 = (A+C), q_1 = (C+D), and q_2 = (B+D).

Maxwell and Pilliner (1968) proposed r_{MP}:

$$r_{\mathrm{MP}} = \frac{2(\mathrm{AD} - \mathrm{BC})}{(\mathrm{p}_1\mathrm{q}_1 + \mathrm{p}_2\mathrm{q}_2)}\ ; \qquad (5)$$

The phi coefficient (as cited in Fleiss, 1975) was proposed as an inter-rater index:

$$\Phi = \frac{(\mathrm{AD} - \mathrm{BC})}{\sqrt{\mathrm{p}_1\mathrm{q}_1\mathrm{p}_2\mathrm{q}_2}}\ ; \qquad (6)$$

The Dice index (as cited in Fleiss, 1975) was proposed as a measure of inter-rater agreement:

$$S_{\mathrm{D}} = \frac{2(\mathrm{AD} - \mathrm{BC})}{(\mathrm{p}_1\mathrm{q}_2 + \mathrm{p}_2\mathrm{q}_1)}\ ; \qquad (7)$$

Methods

The data for the subsequent situations tested on each of the inter-rater agreement indices were derived from an SPSS (Statistical Package for the Social Sciences v. 15.0) program written by the author. Each of the 2 x 2 situations used binary data; there were no missing data; each situation had free, homogeneous or heterogeneous marginals (i.e., "… a margin is 'free' whenever the marginal proportions are not known to the assigner beforehand" (Brennan & Prediger, 1981, p. 690); and each situation had either no rater disagreement, rater disagreement ≤.10, rater disagreement >.10 but ≤.20, or rater disagreement >.20. Rater disagreement was determined from the following formula presented in Sim and Wright (2005):

$$\text{Rater Disagreement} = |\mathrm{B} - \mathrm{C}|\ /\ N \qquad (8)$$

An SPSS bootstrap program created by the author was used. The bootstrap is a resampling method where the sampling properties of a statistic, in this instance the inter-rater agreement indices, are derived by recomputing their value for artificial samples. Thus, the sample data from this study served as pseudo-populations and 20,000 random samples with replacement were drawn from these full samples. Twenty thousand iterations were used as an established threshold where all of the four cases in this study had convergence. Once the bootstrap method was repeated 20,000 times on each of the four cases, a distribution of bootstrapped estimates for the kappa-related indices emerged, where the mean value (i.e., κ_{Boot}) of each bootstrapped distribution was the estimate for each of the four case's population κ value. Further, the bootstrap was employed as a method for estimating generalization error, which, in turn, was used to form 95% confidence intervals around the κ_{Boot}.

Using Monte Carlo generated data, the purpose of this research was to examine the possible effect that equal free row and column marginal proportions (EM), unequal free row and column marginal proportions (UM), and the magnitude of rater disagreement had on the agreement indices under study. As James (1983, p. 651) noted 25 years ago, "Much less attention seems to have been paid to the analysis of nonagreements…"

Results

Table 1 shows the kappa-related statistics in each of the four conditions by sample sizes of 10, 20, 50, and 75 typically found in educational research (Claudy, 1972; Huberty & Mourad, 1980). The bootstrap results from Table 1 denote which of the eight indices were outside of the confidence intervals established as thresholds for each case pertaining to under-estimation or over-estimation of inter-rater agreement given the circumstances of homogeneous or heterogeneous marginals and no rater disagreement to some level of disagreement.

The results found in Table 1 indicated that in the first case, when the marginals were homogeneous (i.e., verified via a McNemar's Test based on difference in the marginal probability distribution between observations in a 2 x 2 matrix, where Ho: p1. = p.1 and H1: p1. ≠ p.1) and there was no rater disagreement present, seven of the eight indices showed no under- or over-estimation of inter-rater agreement, which was an expected assumption in this situation with all of the kappa-like formulas (note: the lone exception of over-estimation was found with the Bennett index).

In the second case, when the marginals were not homogeneous and the amount of rater disagreement was $\leq .10$, there was one instance of over-estimation with the Bennett index and two occurrences of under-estimation found with the Fleiss and Scott indices. In the third case, when the marginals were not homogeneous and the amount of rater disagreement was $> .10$ but $\leq .20$, there were three instance of over-estimation with the Phi, Maxwell-Pilliner, and Armitage et al. indices, and three occurrences of under-estimation found with the Bennett, Fleiss, and Scott indices. Finally, in the fourth case, when the marginals were not homogeneous and the amount of rater disagreement was $> .20$, all of the same indices from case 3 that had over- or under-estimation problems repeated in case 4. That is, there was noticeable over-estimation associated with Phi, Maxwell-Pilliner, and Armitage et al., and evident occurrences of under-estimation found with Bennett, Fleiss, and Scott.

Discussion

Thus, given the similar assumptions affiliated with kappa-like indices of agreement, when there were equal free row and column marginal proportions with no rater disagreement present, seven of the eight indices of agreement yielded the same results. This outcome was expected based on the assumption of marginal homogeneity for many of the kappa-like measures. When there were unequal free row and column marginal proportions and rater disagreement is $\leq .10$, five of the eight indices of agreement tended to produce similar results, with two of the three deviant indices very close to the established confidence interval (e.g., Scott and Fleiss within .001).

When the marginals were not homogeneous and the magnitude of rater disagreement was $> .10$, cases 3 and 4 showed a trend in indices that succumbed to over- and under-estimation. That is, when rater disagreement was evident (i.e., $> .10$), there should be some caution used when applying the Phi, Maxwell-Pilliner, and Armitage et al. indices in a 2 x 2 situation due to their tendency to over-estimate chance-corrected agreement, and some prudence employed when using the Bennett, Fleiss, and Scott indices due to their propensity to under-estimate chance-corrected agreement when compared to other commonly-used indices of agreement.

Implications and Conclusions

Overall, the data trends indicated that the Bennett index either over- or under-estimated chance-corrected agreement in a 2 x 2 situation in all four cases studied regardless of the presence, or lack thereof, of rater disagreement. The Fleiss and Scott indices under-estimated in three of the four cases (i.e., contingent upon some level of rater disagreement).

Table 1. Measures of Agreement Bootstrap Results

Sample	**N = 10**	**N = 20**	**N = 50**	**N = 75**

Cell Counts	A = 2, B = 2,
	C = 2, D = 4
		A = 8, B = 3,
		C = 4, D = 5
			A = 19, B = 4,
			C = 12, D = 15
				A = 30, B = 3,
				C = 21, D = 21

Rater Disagreement	**0**	**≤ .10**	**> .10 ≤ .20**	**> .20**
Agreement Index				
Cohen	.167	.286	.372	.387
Maxwell-Pilliner	.167	.287	.392*	.435*
Scott	.167	.284*	.356*	.351*
Fleiss	.167	.284*	.356*	.351*
Armitage et al.	.167	.287	.392*	.436*
Dice	.167	.286	.372	.387
Phi	.167	.287	.392*	.435*
Bennett	.200*	.300*	.360*	.360*
Bootstrap				
Mean: κ_{Boot}	.171	.288	.374	.393
Standard Deviation	.004	.002	.005	.013
95% Confidence Interval	(.167, .179)	(.285, .292)	(.364, .385)	(.368, .419)

* = Outside of confidence interval range

As seen in Table 2, the Bennett, Scott, and Fleiss indices, which all adhered to the assumption of homogeneous marginals, preformed the poorest when any level of rater disagreement was present and, thus, their use in situations of disagreement is not recommended.

The Phi, Maxwell-Pilliner, and Armitage et al. indices over-estimated in two of the four cases, particularly when rater disagreement > .10, Therefore, the recommendation found in Table 2 is to employ these indices when rater disagreement is ≤ .10. Cohen and Dice were the only indices that did not manifest any penchant to over- or under-estimate chance-corrected agreement when confronted with rater disagreement and are recommended as reliable measures in all conditions tested.

Table 2. Recommendations for the Use of Agreement Indices per Level of Rater Disagreement

Rater Disagreement	**0**	**≤ .10**	**> .10 ≤ .20**	**> .20**
Agreement Index				
Cohen	*	*	**	**
Maxwell-Pilliner	*	*	NR	NR
Scott	*	NR	NR	NR
Fleiss	*	NR	NR	NR
Armitage et al.	*	*	NR	NR
Dice	*	*	**	**
Phi	*	*	NR	NR
Bennett	NR	NR	NR	NR

NR = Not Recommend for use
* = Use in conditions of rater disagreement **≤ .10**
** = Use in conditions of rater disagreement **> .10**

An implication affiliated with the current study may be seen in the area of contributing to the base in the scholarly literature, where this is one of very few studies (cf. Whitehurst, 1984) that has looked at the magnitude that rater disagreement has on various inter-rater agreement indices. As Zwick (1988) noted about the degree that marginal homogeneity may play in inter-rater agreement indices, “Rather than ignoring marginal disagreement or attempting to correct for it, researchers should be studying it to determine whether it reflects important rater differences or merely random error” (p. 377). A second implication is that this research provides guidelines concerning which of the frequently used measures of agreement would be plausible options to employ when a level of rater disagreement is present.

References

Armitage, P., Blendis, L. M., & Smyllie, H. C. (1966). The measurement of observer disagreement in the recording of signs. *Journal of the Royal Statistical Society, Series A, 129*, 98-109.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18*, 303-308.

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

Claudy, J. G. (1972). A comparison of five variable weighting procedures. *Educational and Psychological Measurement, 32*, 311-322.

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

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31*, 651-659.

Hertzberg, V. S., Xu, F., & Haber, M. (2006). Restricted quasi-independent model resolves paradoxical behaviors of Cohen's kappa. *Journal of Modern Applied Statistical Methods*, *5* (2), 417-431.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement, 40*, 101-112.

James, I. R. (1983). Analysis of nonagreements among multiple raters. *Biometrics, 39*, 651-657.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research, 4*, 8-18.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika, 44*, 461-471.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*, 411-433.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice-Hall.

Maxwell, A. E., & Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology, 21*, 105-116.

Rae, G. (1988). The equivalence of multiple rater kappa statistics and intraclass correlation coefficients. *Educational and Psychological Measurement, 48*, 367-374.

Riffe, D., & Freitag, A. A. (1997). A content analysis of content analyses: Twenty-five years of *Journalism Quarterly*. *Journalism & Mass Communication Quarterly, 74*, 873-882.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*, 257-268.

Sirotnik, K. A. (1981). Assessing attitudinal congruence: A case for absolute (as well as relative) indices. *Journal of Educational Measurement, 18*, 205-212.

Übersax, J. S. (1987). Diversity of decision making models and the measurement of interrater agreement. *Psychological Bulletin, 101*, 140-146.

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39*, 22-28.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 374-378.

Send correspondence to: David A. Walker
Northern Illinois University
Email: dawalker@niu.edu