Journal of Modern Applied Statistical Methods

Volume 5 | Issue 2 Article 16

11-1-2005

Restricted Quasi-Independent Model Resolves Paradoxical Behaviors of Cohen's Kappa

Vicki Stover Hertzberg
Emory University, vhertzb@sph.emory.edu

Frank Xu
Emory University, fxu1967@hotmail.com

Michael Haber
Emory University, mhaber@sph.emory.edu

Journal of Modern Applied Statistical Methods
November, 2006, Vol. 5, No. 2, 417- 431

1538 – 9472/06/$95.00

Restricted Quasi-Independent Model Resolves Paradoxical Behaviors of Cohen's Kappa

Vicki Stover Hertzberg Frank Xu Michael Haber
Department of Biostatistics
Emory University

Cohen's kappa, an index of inter-rater agreement, behaves paradoxically in 2×2 tables. λ_A is derived, an index from the restricted quasi-independent model for 2×2 tables. Simulation studies are used to demonstrate λ_A has superior performance compared to Scott's pi. Moreover, λ_A does not show paradoxical behavior for 2×2 tables.

Keywords: Quasi-independent model; Cohen's kappa; Scott's pi; inter-rater agreement

Introduction

In clinical trials and epidemiology studies, agreement studies are often conducted in order to assess and characterize the extent to which two sets of measurements on the same unit of observation agree. Examples of such studies include when raters examine a group of subjects to determine the presence or absence of a trait, sort them into previously arranged categories, or rate them according to a previously defined scale. Examples of areas in which rater variability is of concern include the interpretations of image results in radiology, diagnoses made on the basis of laboratory measurements, or psychiatric evaluations. Data from a study in which raters A and B classify $N_{..}$ subjects into k categories are the counts $\{N_{ij}: i=1,\ldots,k; j=1,\ldots,k\}$ where N_{ij} is the number of subjects that are simultaneously classified as category i by rater A and category j by rater B.

A variety of measures are available to assess the extent of agreement between ratings. Because some agreement can be expected merely due to chance, an important consideration in selecting such a measure is whether or not it is a chance-corrected index. The more popular indices that are chance-corrected include the S statistic (Bennett, et al., 1954), Scott's pi (Scott, 1955), and Cohen's kappa (Cohen, 1960). Among these measures, Cohen's kappa is a popular choice, due to its intuitive means for correcting for chance. The population value for Cohen's kappa can be written as

$$\kappa = \frac{\pi_0 - \pi_e}{1 - \pi_e} \qquad (1)$$

where π_0 is the proportion of observed agreement and π_e is the proportion of agreement expected by chance alone. Cohen's kappa, κ, is estimated in the k×k table as

$$\hat{\kappa} = \frac{\sum_{i=1}^{k} \frac{N_{ii}}{N_{..}} - \sum_{i=1}^{k} \frac{N_{i.}N_{.i}}{N_{..}}}{1 - \sum_{i=1}^{k} \frac{N_{i.}N_{.i}}{N_{..}}} \qquad (2)$$

Vicki Stover Hertzberg has worked in academic biostatistics at University of Cincinnati and Emory University in the Department of Biostatistics. Email: vhertzb@sph.emory.edu. Frank Xu works as a statistician for Spectrx, a medical device manufacturing firm in the Atlanta metropolitan area. Email: fxu1967@hotmail.com. Michael Haber is in the Department of Biostatistics. Email: mhaber@sph.emory.edu. This work was supported in part by US Public Health Service grant 1 R01 MH070028-01A1.

where

$$N_{i.} = \sum_{j=1}^{k} N_{ij} ,$$

$$N_{.j} = \sum_{i=1}^{k} N_{ij} ,$$

and

$$N_{..} = \sum_{i=1}^{k} \sum_{j=1}^{k} N_{ij} .$$

There are a variety of generalizations of Cohen's kappa, such as versions that are weighted for ordinal scale assessments as well as versions for use in the assessment of multi-rater agreement. In this article, the discussion is confined to the assessment of the agreement and disagreement between two raters in a simple square contingency table.

Despite its popularity as an index of agreement, Cohen's kappa exhibits paradoxical behaviors in 2×2 tables (Feinstein & Cicchetti, 1990). For a given 2×2 table, the marginal probabilities are called symmetrical if either ($N_{1.}/N_{..} \geq 0.5$ and $N_{.1}/N_{..} \geq 0.5$) or ($N_{1.}/N_{..} \leq 0.5$ and $N_{.1}/N_{..} \leq 0.5$). The marginal probabilities are called balanced if both $N_{1.}/N_{..}$ and $N_{.1}/N_{..}$ are close to 0.5. One such paradox is that κ estimated for a table with symmetrical unbalanced marginal probabilities can be substantially less than κ estimated for a table with symmetrical balanced marginal probabilities although both tables have the same amount of observed agreement. In addition, a table with asymmetrical unbalanced marginal probabilities will have larger estimated κ than a table with symmetrical unbalanced marginal probabilities even though the observed agreement is the same, the second paradox.

Several authors (Brennan, et al., 1981, Cicchetti & Feinstein, 1990, Lantz & Nebenzahl, 1996, Byrt, et al., 1993) investigated this problematic behavior. They have suggested companion statistics to be reported along with Cohen's kappa; however these companion statistics are not model-based and are arbitrary in the treatment of the correction for chance.

Thus an alternative index which does not exhibit such paradoxical behaviors is desirable. The use of a measure of agreement is explored; λ_A, derived from the quasi-independent (QI) model (Goodman, 1968). The QI model was developed for application to k×k tables, specifically for the analysis of truncated tables (i.e., tables with missing entries due to study design or data collection). One limitation of the QI model is that it is not directly applicable to 2×2 tables. This limitation is due to lack of degrees of freedom. In this article, a restricted QI model for interrater agreement that allows for rater bias in 2×2 tables is examined. The introduction of the restriction allows us to overcome the problem with degrees of freedom.

The notion of quasi-independence assumes that a sub-table, which is part of the whole table, is independent (Bishop, et al., 1975, Agresti, 1990). A two-dimensional table is said to be QI if for a subset of cells U there exist constants p_{ri} and p_{cj} such that the probability of cell (i,j) given it is in U equals $p_{ri}p_{cj}$. The remaining cells are in U*.

Guggermoos-Holzman and Vonk (1998) showed that the QI model is related to latent class models. This relationship is exploited to apply the QI concept to the context of rater agreement studies. Suppose that there are two groups of subjects (latent classes) to be classified into k categories. Group 1 is systematically classified by all raters. If the raters agree on the classification then systematic agreement is said to have occurred; otherwise systematic disagreement has resulted due to the use of different classification rules by the raters.

For Group 1 subjects the classifications by the raters are not made independently, thus they contribute only to U*, the set of cells with systematic agreement or disagreement. Group 2 comprises subjects for whom at least one rater randomly classifies according to a multinomial distribution, that is, the raters classify these subjects with independent marginal probabilities p_{ri} and p_{cj} respectively. Group 2 subjects contribute to the frequencies of all cells in the table. To illustrate this concept, consider Table 1. In this scenario, raters A and B classify 100 subjects into three categories. Unbeknownst to A and B there are 80 subjects in Group 1 and 20 in

Table 1. Illustrative Example of Quasi-independent Data.

Easy to Classify Subjects (Group 1)						Difficult to Classify Subjects (Group 2)					Whole Table			
	Rater A category					Rater A category					Rater A category			
Rater B category	1	2	3	Total		1	2	3	Total		1	2	3	Total
1	**25**[1]	**5**		30		4	4	2	10		**29**	**9**	2	40
2		**30**		30	+	2	2	1	5	=	2	**32**	1	35
3			**20**	20		2	2	1	5		2	2	**21**	25
Total	25	35	20	80		8	8	4	20		33	43	24	100

[1]Cells in set U* denoted with boldface.

Group 2. Group 2 classifications are made using independent marginal probabilities of (0.5, 0.25, 0.25) for categories 1, 2, and 3 respectively by rater A and (0.4, 0.4, and 0.2) by rater B. The set U* comprises cells (1,1), (1,2), (2,2), and (3,3). In U*, the cells (1,1), (2,2), and (3,3) represent systematic agreement, while the cell (1,2) represents systematic disagreement. Systematic disagreement may arise when the raters use slightly different rules for classification. In the case of Table 1, the rules used by rater A are such that s/he tends to over-read category 2 subjects versus category 1 in comparison to rater B.

Suppose that λ is the proportion of the population of subjects in Group 1 and 1-λ is the proportion in Group 2. Thus λ is the total proportion of systematic agreement and disagreement. If cell (i,j) is in U* then define d_{ij} = 1, and d_{ij} = 0 otherwise. When i=j, χ_{ij} is the proportion of systematic agreement and when i≠j, χ_{ij} is the proportion of systematic disagreement, defined only for cells in U*. For each cell (i,j) let $\frac{d_{ij}\chi_{ij}}{\lambda}$ be the conditional probability that a subject is classified into that cell given that it is in Group 1. Thus $\lambda = \sum_{i=1}^{k}\sum_{j=1}^{k} d_{ij}\chi_{ij}$. By the total probability theorem, the probability of cell (i,j) can be written as

$$\pi_{ij} = (1-\lambda)p_{r_i}p_{c_j} + d_{ij}\chi_{ij}. \qquad (3)$$

One may solve for λ by multiplying both sides of equation (3) by d_{ij} and summing over all cells obtaining

$$\lambda = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} d_{ij}\pi_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} d_{ij}p_{r_i}p_{c_j}}{1-\sum_{i=1}^{k}\sum_{j=1}^{k} d_{ij}p_{r_i}p_{c_j}}. \qquad (4)$$

Note the similarity of λ to the formulation of a chance-corrected agreement index. In this formulation the terms d_{ij} are terms that must be specified before any further calculations can be made. There are k^2-1 degrees of freedom available in the k×k table, of which 2(k-1) are the parameters for the marginal probabilities. Thus, at most $(k-1)^2$ parameters of the d_{ij} can be set to 1 in equation (4). As a result, the QI model can only be used in k×k tables where k ≥ 3.

Furthermore, λ can be expressed as the sum of systematic agreement and disagreement as follows:

$$\lambda = \lambda_A + \lambda_D = \sum_{i=1}^{k} d_{ii}\chi_{ii} + \sum_{i=1}^{k}\sum_{j\neq i} d_{ij}\chi_{ij} \ . \tag{5}$$

The log-likelihood of the general QI model then is given by

$$\ln(l) = \sum_{i=1}^{k}\sum_{j=1}^{k} N_{ij} \ln\left[(1-\lambda)p_{r_i}p_{c_j} + d_{ij}\chi_{ij}\right]. \tag{6}$$

The unknown parameters are p_{ri}, p_{cj}, and χ_{ij}, with $\lambda=\sum\chi_{ij}$.

The following iterative procedure may be used to derive the maximum likelihood estimates for the model:

$$\hat{p}_{r_i}^{(l)} = \frac{p_{i\cdot} - \sum_{j=1}^{k} d_{ij}\hat{\chi}_{ij}^{(l)}}{1-\hat{\lambda}^{(l)}}$$

$$\hat{p}_{c_j}^{(l)} = \frac{p_{\cdot j} - \sum_{i=1}^{k} d_{ij}\hat{\chi}_{ij}^{(l)}}{1-\hat{\lambda}^{(l)}} \tag{7}$$

where

$$\hat{\chi}_{ij}^{(l)} = d_{ij}\left[p_{ij} - (1-\hat{\lambda}^{(l)})\hat{p}_{r_i}^{(l)}\hat{p}_{c_j}^{(l)}\right],$$

$$\hat{\lambda}^{(l)} = \sum_{i=1}^{k}\sum_{j=1}^{k} d_{ij}\hat{\chi}_{ij}^{(l)}$$

and $p_{i\cdot}$ and $p_{\cdot j}$ are the observed marginal probabilities. Initial values are $\hat{\lambda}^{(0)}$=0, and $\hat{p}_{r_i}^{(0)}$ and $\hat{p}_{c_j}^{(0)}$ are set to the observed marginal probabilities, i, j = 1,…,k.

To derive these estimates one must set d_{ij} = 1 or 0 on the basis of either *a priori* knowledge or using a data driven method. Agresti (1990) assumed d_{ij} = 1 for all diagonal cells while Bergan (1980) and Aickin (1990) used a trial and error method to determine d_{ij} from the data.

Some illustrations are explored. Returning to Table 1, it may be seen that λ=0.8. The values of χ_{ij} are 0.25, 0.05, 0.30, and 0.20 for cells in U* and not defined otherwise.

Next, turn to Table 2. If systematic agreement is assumed in every diagonal cell and no systematic disagreement, then $\hat{\lambda}$(se) = $\hat{\lambda}_A$(se) = 0.554(0.008), compared to $\hat{\kappa}$(se) = 0.493(0.057), where standard errors are obtained by bootstrap. A goodness of fit test for this table results in χ^2 = 11.7 with 5 degrees of freedom, p=0.039, giving an indication of lack of fit.

Table 2. Diagnosis of Carcinoma for Pathologists A and B

Classification of Pathologist A	Classification of Pathologist B				
	1	2	3	4 & 5	Total
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4 & 5	0	1	17	10	28
Total	27	12	69	10	118

Source: Derived from Landis and Koch (1977) as described in Agresti (1990)

Note the relatively small amount of agreement in cell (2,2) (5.9% of the 118 observations versus 18.6%, 30.5%, and 8.5% in cells (1,1), (3,3), and (4,4) respectively) and the large error frequency in cell (4,3) (14.4%). Setting U* to contain cells (1,1), (3,3), (4,4), and (4,3), the following are obtained $\hat{\lambda}$(se) = 0.69(0.005), with $\hat{\lambda}_A$(se) = 0.554(.005) and $\hat{\lambda}_D$(se) = 0.136(.003). χ^2 = 2.18 with 5 degrees of freedom, p=0.82, may be further derived from this model, indicating much better fit.

Methodology

For the case of 2×2 tables, it is assumed that i=1 or j=1 indicates that the prevalent condition is positive. Due to lack of degrees of freedom, it must also be assumed in this case that only agreement is systematic, i.e., there is no systematic disagreement. Thus U* contains only the two diagonal cells. The model can now be rewritten as

$$\pi_{ij} = (1-\lambda_A)p_{r_i}p_{c_j} + I_{(i=j)}\chi_{ij} \qquad (8)$$

for i = 1, 2 and j = 1, 2, where $0 \le p_{ri} \le 1$, $p_{r1} + p_{r2} = 1$, $0 \le p_{cj} \le 1$, $p_{c1} + p_{c2} = 1$, $0 \le \chi_{ii} \le 1$, and $0 \le \lambda = \chi_{11} + \chi_{22} \le 1$.

As mentioned above, there are three degrees of freedom and four parameters: χ_{11}, χ_{22}, p_{r1}, and p_{c1}. If no restriction is placed on the independent marginal probabilities then a restriction must be placed on χ_{11} and χ_{22}.

The common correlation model for Scott's pi, denoted κ_s, assumes 1) no rater bias and 2) the rater prevalence in Group 1 equals that in Group 2. The second assumption follows from 3) the underlying true prevalence in Group 1 equals that in Group 2 and 4) the common rater prevalence is an unbiased estimator of the true prevalence. Scott's pi is limited by the assumption of no rater bias, in particular the assumption of no rater bias in Group 2. It is theorized that the two observers are likely to have different rater prevalence's in Group 2. In fact, many agreement studies show evidence of rater bias. Thus, to adequately apply the QI concept to rater agreement in 2×2 tables, one must assume (5) the true prevalence in Group 2 is equal to that in Group 1, and (6) the rater prevalence in Group 2 differs between raters, but the average is an unbiased estimator for the true prevalence.

Under assumption (6), χ_{11}/λ_A can be interpreted as the common rater prevalence in Group 1 since the two raters classify with certainty and agree. Thus, χ_{11}/λ_A is the best estimator of true prevalence. If then one allows for the prevalence for each rater for Group 2 subjects, then $(p_{r1}+p_{c1})/2$ is also an estimator of true prevalence. Under assumption (5) then one has $\chi_{11}/\lambda_A=(p_{r1}+p_{c1})/2$, giving $\chi_{ii}=(p_{ri}+p_{ci})\lambda_A/2$. Thus, under assumptions (5) and (6) one has the restricted QI model,

$$\pi_{ij} = (1-\lambda_A)p_{r_i}p_{c_j} + I_{(i=j)}\frac{p_{ri}+p_{cj}}{2}\lambda_A \qquad (9)$$

for i = 1, 2 and j = 1, 2 where $0 \le p_{ri} \le 1$, $p_{r1} + p_{r2} = 1$, $0 \le p_{cj} \le 1$, $p_{c1} + p_{c2} = 1$, $0 \le \chi_{ii} \le 1$, and $0 \le \lambda = \chi_{11} + \chi_{22} \le 1$. The log-likelihood function for this restricted model is

$$\ln(l) = \sum_{i=1}^{2}\sum_{j=1}^{2} N_{ij} \ln\left[\begin{array}{l}(1-\lambda_A)p_{r_i}p_{c_j} \\ +I_{(i=j)}\dfrac{p_{ri}+p_{cj}}{2}\lambda_A\end{array}\right]. \qquad (10)$$

The maximum likelihood estimators can be derived by setting the score equations with respect to the parameters equal to zero and solving for the unknown parameters. Alternatively, estimates can be obtained by solving the equations $E(N_{ij}) = N_{ij}$, i = 1, 2, j = 1, 2.

Thus, the following maximum likelihood estimates are obtained:

$$\hat{\lambda}_A = \frac{N_{11}}{2N_{11}+N_{12}+N_{21}} + \frac{N_{22}}{2N_{22}+N_{12}+N_{21}} - \frac{T}{(2N_{11}+N_{12}+N_{21})(2N_{22}+N_{12}+N_{21})} \qquad (11)$$

where

$$T=\sqrt{\begin{array}{l}(N_{11}-N_{22})^2(N_{12}+N_{21})^2\\+4N_{21}N_{12}(2N_{11}+N_{12}+N_{21})\\(2N_{22}+N_{12}+N_{21})\end{array}}$$

$$\hat{p}_{r1}=\frac{N_{11}+N_{12}}{N_{..}}-\frac{(N_{12}-N_{21})\hat{\lambda}_A}{2N_{..}(1-\hat{\lambda}_A)}$$

and

$$\hat{p}_{c1}=\frac{N_{11}+N_{21}}{N_{..}}-\frac{(N_{12}-N_{21})\hat{\lambda}_A}{2N_{..}(1-\hat{\lambda}_A)}$$

Note that when $N_{11}N_{22}=N_{12}N_{21}$ (independence), $\hat{\lambda}_A=0$ and when $N_{12}=N_{21}=0$ (total agreement), $\hat{\lambda}_A=1$. Moreover, the estimated independent marginal probabilities are the same as the observed marginal probabilities iff $\hat{\lambda}_A=0$ or $N_{12}=N_{21}$.

If both N_{12} and N_{21} are replaced by $(N_{12}+N_{21})/2$ on RHS of equation (11) the estimator for Scott's pi, κ_S is derived. When the assumption of no rater bias is made, the extended model (9) reduces to the common correlation model (Donner & Eliasziw, 1992). Because the extended model allows for rater bias, $\hat{\lambda}_A$ has several advantages over κ_S as follows:

1. The common correlation model does not fit the data if in fact there is substantial rater bias. Thus applications of the common correlation model of Scott's pi could be misleading in the absence of a test for rater bias, such as McNemar's test. However, even with such a test, the power may be insufficient to detect rater bias for a small sample size, and an improper application of Scott's pi may occur.
2. Both the common correlation model and the extended model assume only random agreement, i.e., no systematic disagreement. If the two raters have different independent marginal probabilities they are less likely to agree with each other by chance, and the observed agreement is more likely to have been achieved for reasons other than chance. Therefore, if there is rater bias, the estimated agreement index should increase with this bias given the same amount of observed agreement. Scott's pi does not change with observer bias while $\hat{\lambda}_A$ increases with increasing rater bias.
3. When the table is independent ($N_{11}N_{22}=N_{12}N_{21}$) one would expect the estimated systematic agreement to be zero, as is the case with $\hat{\lambda}_A$. However Scott's pi does not equal zero unless $N_{12}=N_{21}$ in addition.

It may be written

$$\kappa_S=\frac{(1-\lambda_A)\sum_{i=1}^{2}p_{ri}p_{ci}+\lambda_A-\sum_{i=1}^{2}\left(\frac{p_{ri}+p_{ci}}{2}\right)^2}{1-\sum_{i=1}^{2}\left(\frac{p_{ri}+p_{ci}}{2}\right)^2} \tag{12}$$

and

$$error(\kappa_s,\lambda_A)=\kappa_s-\lambda_A=\frac{(1-\lambda_A)(p_{r1}-p_{c1})^2}{2(1-\sum_{i=1}^{2}\left(\frac{p_{ri}+p_{ci}}{2}\right)^2)} \tag{13}$$

When there is no rater bias, $p_{r1}=p_{c1}$ and $\kappa_s=\lambda_A$. As p_{r1}-p_{c1} increases, so does $error(\kappa_s,\lambda_A)$. Table 3 shows the values of $error(\kappa_s, \lambda_A)$ at different values of $(p_{r1},p_{c1},\lambda_A)$.

Table 3. Error of Scott's pi under the Restricted QI Model

p_{r1}	p_{c1}	λ =0	λ=0.1	λ=0.3	λ=0.5	λ=0.7	λ=0.9
0.9	0.1	0.640	0.576	0.448	0.320	0.192	0.064
0.8	0.2	0.360	0.324	0.252	0.180	0.108	0.036
0.7	0.3	0.160	0.144	0.112	0.080	0.048	0.016
0.6	0.4	0.040	0.036	0.028	0.020	0.012	0.004

Table 4. Diagnosis of Carcinoma for Pathologists A and B

	Classification of Pathologist B		Total
Classification of Pathologist A	Class 1 or 2	Class 3 or 4 or 5	
Class 1 or 2	36	16	52
Class 3 or 4 or 5	3	63	66
Total	39	79	118

Consider Table 4 which shows the collapsed version of Table 2. The p-value for McNemar's test for this table is 0.003, indicating substantial rater bias. For this table, $\hat{\kappa}_S$=0.660 and $\hat{\lambda}_A$=0.703 are obtained.

Results

It has been shown above that $\hat{\lambda}_A$ increases with increasing rater bias while Scott's pi does not change, and that $\hat{\lambda}_A$= 0 for an independent 2×2 table, whereas Scott's pi is only zero when the off-diagonal elements are equal to each other. Thus, λ_A is preferable to κ_S as a measure of agreement in the presence of rater bias. When there is no rater bias, is λ_A as good as κ_S? Simulations were conducted to investigate the performance of these agreement indices.

Simulations were conducted for a total of 100 configurations: 4 different sample sizes, $N_{..}$={20,50,100,200}; 5 different nominal values of systematic agreement λ_A = {0,0.1,0.3,0.6,0.9}; and 5 combinations of independent marginal probabilities (p_{r1},p_{c1})={(0.9,0.9),(0.8,0.8),(0.7,0.7),(0.6,0.6),(0.5,0.5)}. The four values for sample size range from small to moderate to sufficiently large. The five values of λ_A cover the whole range. The independent marginal probabilities all represent the case where there is no rater bias. Because of symmetry, only probabilities of 0.5 to 0.9 are investigated.

The probability of each cell is computed according to the extended model specified in (8) given the nominal values of systematic agreement and independent marginal probabilities. The frequency of each cell is generated as a multinomial random number

given the sample size, using the GENMUL routine (Brown and Lovato). For each configuration, 1000 tables were generated.

The efficiency of the two indices were compared in terms of empirical bias, empirical standard deviation (defined as the standard deviation of the estimated values from the 1000 tables) and empirical residual mean square error (RMSE) (defined as the square root of the mean square of differences between the estimated values of the index and the nominal values over the 100 tables). Figure1 displays a side-by-side comparison of the bias of $\hat{\lambda}_A$ compared to the bias of $\hat{\kappa}_S$ as a function of the nominal values of systematic agreement, independent marginal probability and sample size. Figures 2 and 3 give similar displays for the standard deviations and RMSEs.

From these figures, it is observed that $\hat{\kappa}_S$ is negatively biased, underestimating the true value of systematic agreement in all situations, whereas $\hat{\lambda}_A$ is either positively or negatively biased. There is an increasing trend in bias for $\hat{\lambda}_A$ as λ_A increases. Both indices are increasingly likely to underestimate λ_A as $p_{r1}=p_{c1}$ increases. The biases of both $\hat{\lambda}_A$ and $\hat{\kappa}_S$ decrease as sample size increases. There are no differences between the standard errors and RMSE values of $\hat{\lambda}_A$ and $\hat{\kappa}_S$ that are visually discernible. The standard error and RMSE of both indices tend to be smaller when λ_A is close to either 0 or 1, when the independent marginal probability is close to 0.5, or when the sample size is large.

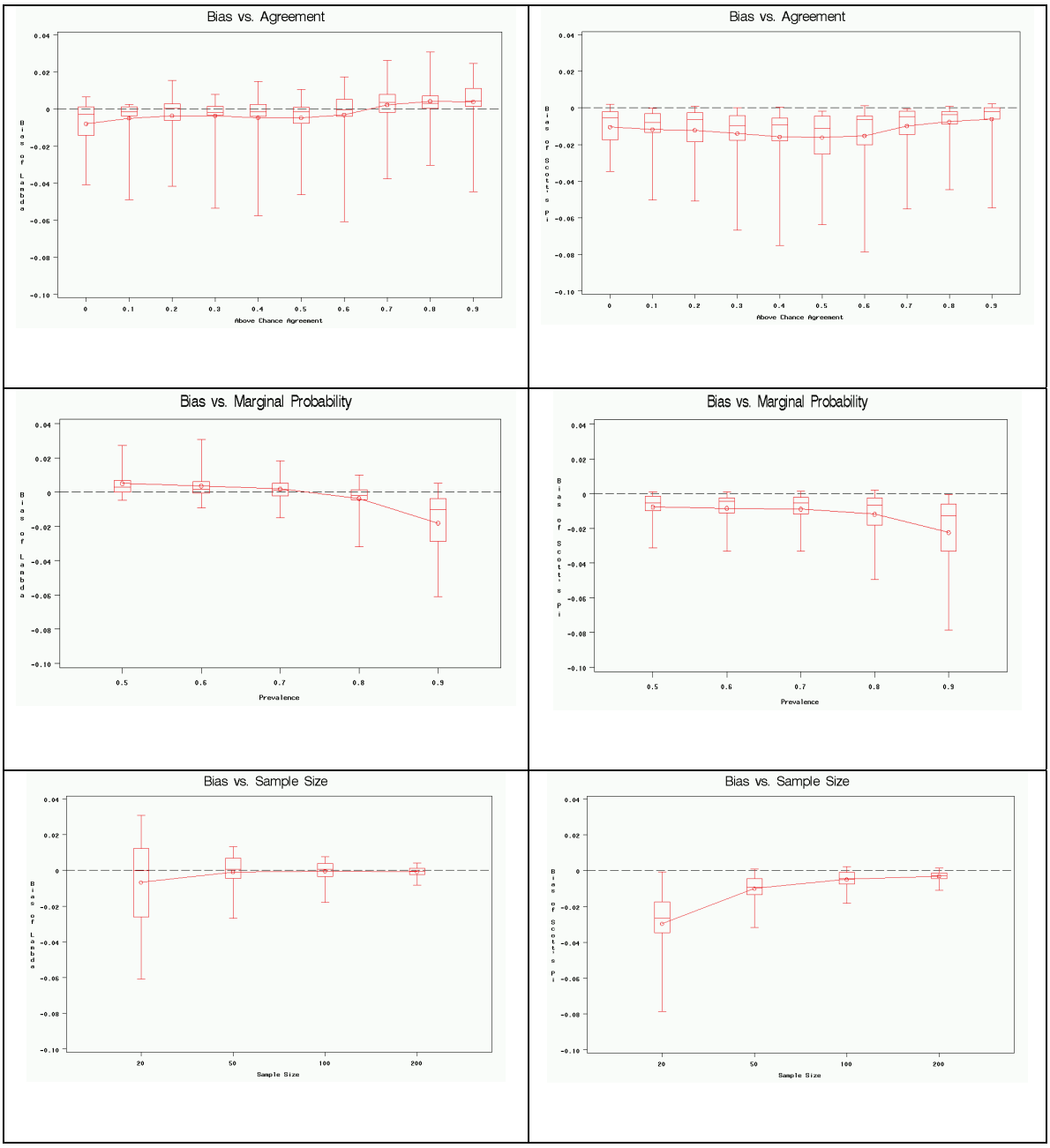

Figure 1. Side-by-side comparison of bias of $\hat{\lambda}_A$ (Lambda) and $\hat{\kappa}_S$ (Scott's Pi) as a function of nominal values for systematic agreement, independent marginal probability, and sample size.

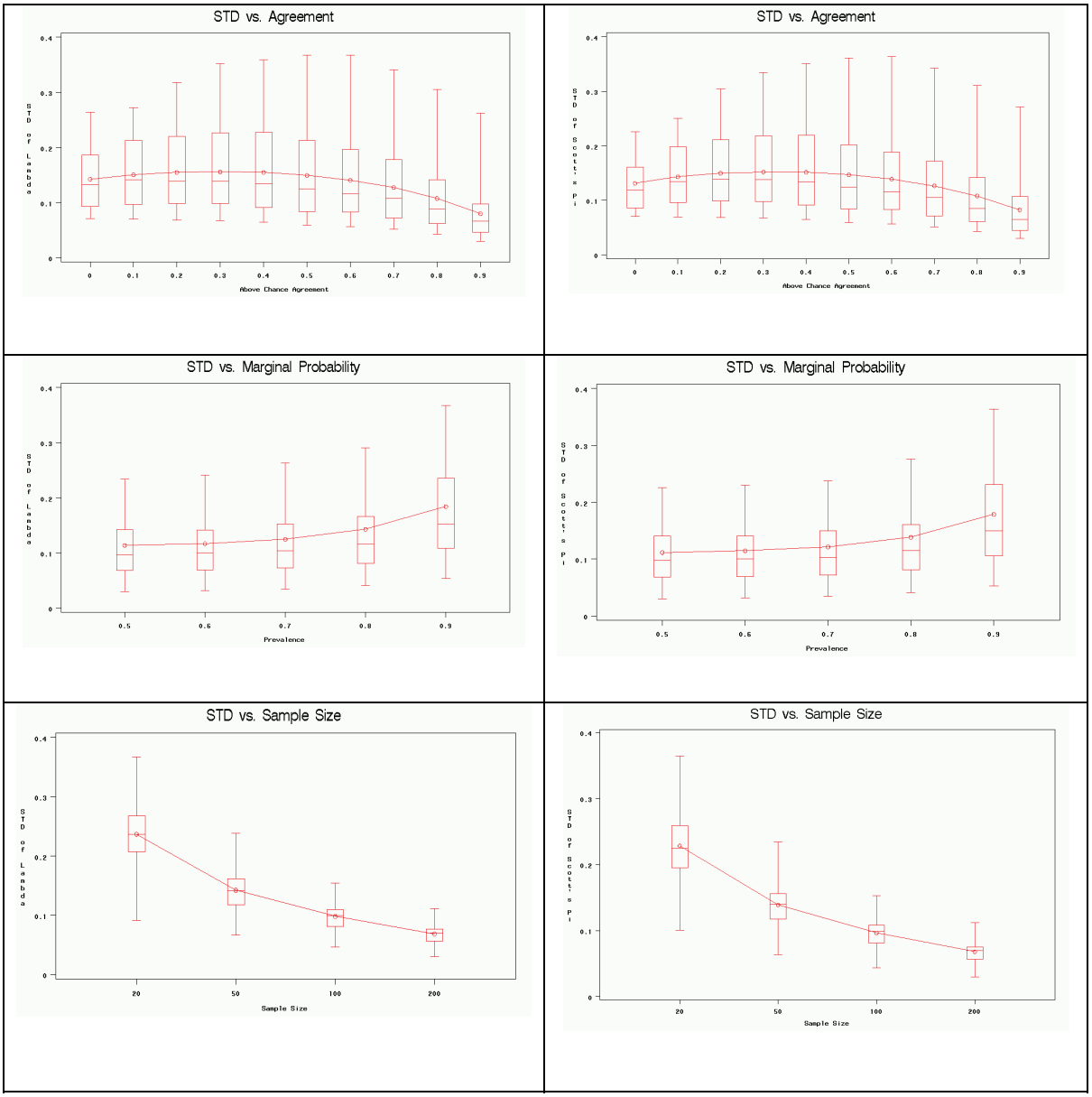

Figure 2 Side-by-side comparison of standard deviation (STD) of $\hat{\lambda}_A$ (Lambda) and $\hat{\kappa}_S$ (Scott's Pi) as a function of nominal values for systematic agreement, independent marginal probability, and sample size.

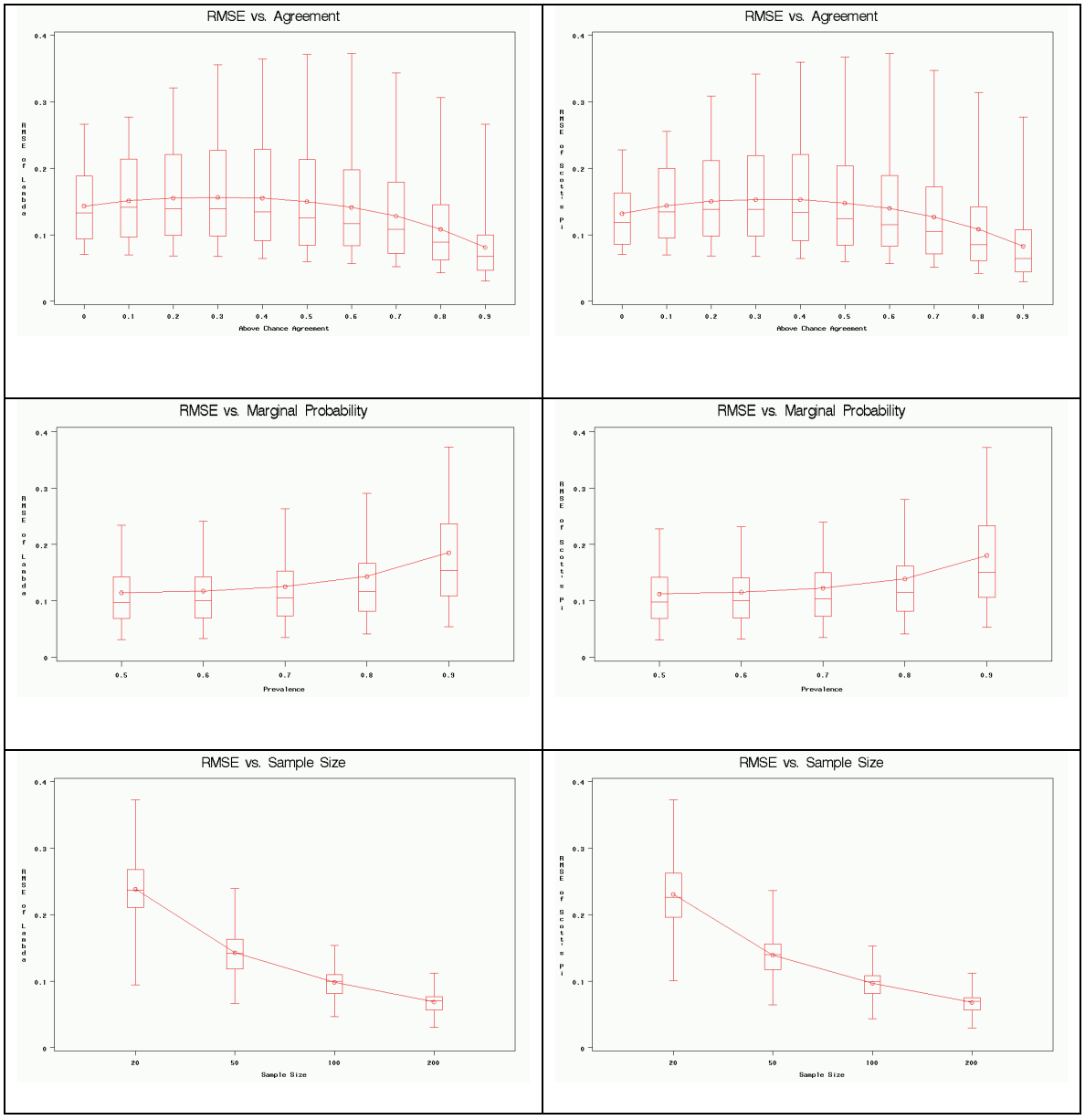

Figure 3. Side-by-side comparison of residual mean square error (RMSE) of $\hat{\lambda}_A$ (Lambda) and $\hat{\kappa}_S$ (Scott's Pi) as a function of nominal values for systematic agreement, independent marginal probability, and sample size.

The new index is now compared to Cohen's kappa. Cohen's kappa corrects for chance agreement using the assumption of independence between raters. However, the assumption of independence in agreement studies is not valid. Some degree of agreement is usually expected in most agreement studies. If there is systematic agreement present, the classifications of the raters cannot be independent since the raters are dealing with the same information (i.e., the same subject) on which to base each of their classifications. The assumption of blindness of ratings is reasonable, leading to an assumption of conditional independence. Thus the formula

$$\Pr(i, j) = p_{r_i} p_{c_j} \qquad (14)$$

should not be used to estimate the expected agreement by chance.

Alternatively, one can use the QI concept to investigate agreement in this context. If it is assumed that no systematic disagreement is present, then equation (3) reduces to

$$\pi_{ij} = (1 - \lambda_A) p_{r_i} p_{c_j} + d_{ii} \chi_{ii} . \qquad (15)$$

Because $\sum_{i=1}^{k} \hat{\pi}_{ii} = p_0$ where p_0 is the observed agreement, an estimator is obtained

$$\hat{\lambda}_A = \frac{\left[p_0 - \sum_{i=1}^{k} \hat{p}_{r_i} \hat{p}_{c_j} \right]}{\left[1 - \sum_{i=1}^{k} \hat{p}_{r_i} \hat{p}_{c_j} \right]} \qquad (16)$$

where $\hat{p}_{r_i}$ and $\hat{p}_{c_j}$ are estimates of the independent marginal probabilities, estimated only in the difficult to classify group (Group 2) where the observers classify independently. Note the similar formulation of $\hat{\lambda}_A$ and $\hat{\kappa}$ given in (2), with the difference being in the marginal probabilities used in calculating these two agreement indices.

If one assumes that there is either no systematic agreement or systematic disagreement, the quasi independent concept resolves the paradoxes posed by Cohen's kappa in 2×2 tables described earlier. Consider the following illustrations.

A. The case of the first paradox is illustrated by the two independent tables shown in Table 5. The values of lambda are 0 for both table 5.1 and table 5.2. All agreement is random agreement. In these cases the observed marginal probabilities are equal to the independent marginal probabilities. Table 5.1 has a set of symmetrical balanced independent marginal probabilities and table 5.2 has a set of symmetrical unbalanced marginal probabilities. Intuitively, the table with unbalanced independent marginal probabilities yields more agreement. A subject in table 5.1 has a 50% chance to be agreed upon by the two observers while the chance is 82% that a subject in table 5.2 will be agreed upon. The agreement is not systematic and can be considered as random agreement. Thus, a set of symmetrical unbalanced marginal probabilities yields more random agreement and less systematic agreement than a set of symmetrical balanced marginal probabilities.

Next consider Table 6. $\hat{\lambda}_A$=0.70 and $\hat{\lambda}_A$=0.32 are calculated for table 6.1 and table 6.2, respectively, which agree with the Cohen's kappa estimates. However, one is also able to use the QI concept to derive estimates of the independent marginal probabilities, obtaining ($\hat{p}_{r_1}$, $\hat{p}_{c_1}$) = (0.53, 0.42) and ($\hat{p}_{r_1}$, $\hat{p}_{c_1}$) = (0.91, 0.84). Thus, table 6.1 yields a set of symmetrical balanced independent marginal probabilities while table 6.2 yields a set of symmetrical unbalanced independent marginal probabilities. Given the same observed agreement, the amount of systematic agreement (estimated by $\hat{\lambda}_A$) should be greater for the tables with symmetrical balanced independent marginal probabilities. Thus, by using the construct of a latent class of subjects who are systematically classified, one arrives at a resolution of the first paradox.

Table 5. Balanced and unbalanced independent marginal probabilities

Table 5.1: $\hat{P}_o$ = 0.50, $\hat{\lambda}_A$=0			
	Observer A		
Observer B	Yes	No	Total
Yes	25	25	50
No	25	25	50
Total	50	50	100

Table 5.2: $\hat{P}_o$ = 0.82, $\hat{\lambda}_A$=0			
	Observer A		
Observer B	Yes	No	Total
Yes	81	9	90
No	9	1	10
Total	90	10	100

Table 6. Balanced and Unbalanced Marginal Probabilities

Table 6.1: Balanced Marginal Probabilities			
	Observer A		
Observer B	Yes	No	Total
Yes	40	9	49
No	6	45	51
Total	46	54	100

Table 6.2: Unbalanced Marginal Probabilities			
	Observer A		
Observer B	Yes	No	Total
Yes	80	10	90
No	5	5	10
Total	85	15	100

For table 6.1 we obtain $\hat{P}_o$ = 0.85, $\hat{\kappa}$ = 0.70, $\hat{\lambda}_A$=0.70, and ($\hat{p}_{r_1}$, $\hat{p}_{c_1}$) = (0.53, 0.42), while for table 6.2 we obtain $\hat{P}_o$ = 0.85, $\hat{\kappa}$ = 0.32, $\hat{\lambda}_A$=0.32, and ($\hat{p}_{r_1}$, $\hat{p}_{c_1}$) = (0.91, 0.84).
Source: Derived from Feinstein and Cicchetti (1990)

B. Consider now the case of the second paradox illustrated by the two independent tables in Table 7. Table 7.1 has a set of symmetrical marginal probabilities while table 7.2 has a set of asymmetrical marginal probabilities. There is no systematic agreement in either table. The observed marginal probabilities are the independent marginal probabilities. Intuitively, the agreement achieved by the observers in Table 7.1 is much more than that in table 7.2 Thus a set of symmetrical independent marginal probabilities yields more agreement than a set of asymmetrical independent marginal probabilities.

Next consider Table 8. $\hat{\lambda}_A$=0.13 and ($\hat{p}_{r_1}$, $\hat{p}_{c_1}$) = (0.59, 0.71), $\hat{\lambda}_A$=0.33 and ($\hat{p}_{r_1}$, $\hat{p}_{c_1}$) = (0.67, 0.23) are estimated for tables 8.1 and 8.2 respectively. The independent marginal probabilities have more symmetry for Table 8.1, yielding more random agreement and less systematic agreement than Table 8.2. Given the same amount of observed agreement, there is less systematic agreement (estimated by $\hat{\lambda}_A$) for Table 8.1, thus resolving the second paradox.

Table 7. Symmetrical and asymmetrical independent marginal probabilities

Table 7.1: $\hat{P}_o$ = 0.82, $\hat{\lambda}_A$=0			
	Observer A		
Observer B	Yes	No	Total
Yes	81	9	90
No	9	1	10
Total	90	10	100

Table 7.2: $\hat{P}_o$ = 0.18, $\hat{\lambda}_A$=0			
	Observer A		
Observer B	Yes	No	Total
Yes	9	81	90
No	1	9	10
Total	10	90	100

Table 8. Symmetrical and Asymmetrical Unbalanced Marginal Probabilities

Table 8.1: Symmetrical Unbalanced Marginal Probabilities			
	Observer A		
Observer B	Yes	No	Total
Yes	45	15	60
No	25	15	40
Total	70	30	100

Table 8.2: Asymmetrical Unbalanced Marginal Probabilities			
	Observer A		
Observer B	Yes	No	Total
Yes	25	35	60
No	5	35	40
Total	30	70	100

For table 8.1 we obtain $\hat{P}_o$ = 0.60, $\hat{\kappa}$ = 0.13, $\hat{\lambda}_A$=0.13 and ($\hat{p}_{r_1}$, $\hat{p}_{c_1}$) = (0.59, 0.71), while for table 8.2 we obtain $\hat{P}_o$ = 0.60, $\hat{\kappa}$ = 0.26, $\hat{\lambda}_A$=0.33 and ($\hat{p}_{r_1}$, $\hat{p}_{c_1}$) = (0.67, 0.23).
Source: Derived from Feinstein and Cicchetti (1990)

Conclusion

It has been shown how the quasi-independent concept can be applied to studies of inter-rater agreement. When applied to 2×2 tables, the use of the QI concept results in a paradigm for agreement that resolves the paradoxical behavior of the popular measure, Cohen's kappa, although the resulting measure can only be derived after the user decides on cells representing systematic agreement or disagreement. This measure has other desirable properties; specifically it allows for assessment of the independent marginal probabilities, which can be reported as companion statistics. Unlike the other statistics that have been suggested for reporting along with Cohen's kappa, these independent marginal probabilities are model-based. Thus, further use and study of the application of the QI concept in inter-rater agreement studies is warranted.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York, N.Y.: John Wiley & Sons.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics, 46*, 293-302.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). The measurement of observer disagreement in the recording of signs. *Journal of the Royal Statistical Society A, 129*, 98-109.

Bergan, J. R. (1980). Measuring observer agreement using the quasi-independent concept. *Journal of Educational Measurement, 17*, 59-69.

Bishop, Y. N., Feinberg, S. W., & Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-698.

Brown, B. W., & Lovato, J. GENMUL routine. Library of C Routines for Random Number Generation. Department of Biomathematics, Box 237, University of Texas, M. D. Anderson Cancer Center, Houston, TX 77030.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology, 46*, 423-429.

Cicchetti, D. V. & Feinstein, A. (1990). High agreement by 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.

Donner, A., & Eliasziw, M. (1992). A goodness-of-fit approach to inference procedure for the kappa statistic: Confidence interval construction, significance testing and sample size estimation. *Statistics in Medicine, 11*, 1511-1519.

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

Goodman, L. A. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing data entries. *Journal of the American Statistical Association, 63*, 1091-1131.

Guggenmoos-Holzmann, I., & Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class prespective. *Statistics in Medicine, 17*, 797-812.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-74.

Lantz, C. A., & Nebenzahl, E. (1996). Behavior and interpretation of the κ statistic; resolution of the two paradoxes. *Journal of Clinical Epidemiology, 49*, 431-434.

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