Journal of Biopharmaceutical Statistics, 17: 629–652, 2007
Copyright © Taylor & Francis Group, LLC
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400701376498

Taylor & Francis
Taylor & Francis Group

A UNIFIED APPROACH FOR ASSESSING AGREEMENT FOR CONTINUOUS AND CATEGORICAL DATA

Lawrence Lin
Baxter Healthcare Co., Round Lake, Illinois and Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, Illinois, USA

A. S. Hedayat
Department of Mathematics, Statistics, and Computer Science, University of Illinois, Chicago, Illinois, USA

Wenting Wu
Division of Biostatistics, Mayo Clinic, Rochester, Minnesota, USA

This paper proposes several Concordance Correlation Coefficient (CCC) indices to measure the agreement among k raters, with each rater having multiple (m) readings from each of the n subjects for continuous and categorical data. In addition, for normal data, this paper also proposes the coverage probability (CP) and total deviation index (TDI). Those indices are used to measure intra, inter and total agreement among all raters. Intra-rater indices are used to measure the agreement among the multiple readings from the same rater. Inter-rater indices are used to measure the agreement among different raters based on the average of multiple readings. Total-rater indices are used to measure the agreement among different raters based on individual readings. In addition to the agreement, the paper also assess intra, inter, and total precision and accuracy. Through a two-way mixed model, all CCC, precision and accuracy, TDI, and CP indices are expressed as functions of variance components, and GEE method is used to obtain the estimates and perform inferences for all the functions of variance components. Each of previous proposed approaches for assessing agreement becomes one of the special case of the proposed approach. For continuous data, when m approaches ∞, the proposed estimates reduce to the agreement indices proposed by Barnhart et al. (2005). When m = 1, the proposed estimate reduces to the ICC proposed by Carrasco and Jover (2003). When m = 1, the proposed estimate also reduces to the OCCC proposed by Lin (1989), King and Chinchilli (2001a) and Barnhart et al. (2002). When m = 1 and k = 2, the proposed estimate reduces to the original CCC proposed by Lin (1989). For categorical data, when k = 2 and m = 1, the proposed estimate and its associated inference reduce to the kappa for binary data and weighted kappa with squared weight for ordinal data.

Key Words: Accuracy; CCC; CP; ICC; Inter-agreement; Intra-agreement; Kappa; MSD; Precision; TDI; Total-agreement.

Received September 10, 2006; Accepted February 8, 2006
Address correspondence to Lawrence Lin, Baxter Healthcare Co., WH2-35, Rt. 120 and Wilson Road, Round Lake, IL 70073, USA; E-mail: Lawrence_Lin@baxter.com

1. INTRODUCTION

Measuring agreements between different methods or different raters have received a great deal of attention recently. Cohen (1960, 1968) and Fleiss et al. (1969), Fleiss and Cohen (1973) proposed kappa and weighted kappa to measure agreement for binary or ordinal data. Lin (1989, 1992, 2000, 2003) proposed the Concordance Correlation Coefficient (CCC), the Total Deviation Index (TDI) and Lin et al. (2002) proposed the Coverage Probability (CP) to measure agreement for continuous data. Lin defined the CCC to be the product of precision and accuracy, which is intuitive and very easy to understand. Both Cohen and Lin considered the case of measuring agreement between two raters, with each rater measures each of the n subjects once. Robieson (1999) proved that the CCC equals to kappa for binary data and weighted kappa with the squared weight set for ordinal data. Generalized Estimating Equations (GEE) approach was introduced to agreement assessment by several authors starting in 2000: Williamson et al. (2000) proposed modelling kappa by GEE approach for categorical data. Barnhart et al. (2002, 2005) proposed modelling CCC by GEE approach for continuous data. The advantages about GEE approach are: a) the pairwise agreements (kappa or CCC) among k raters can be modelled with covariates adjusted. b) this approach doesn't require the full knowledge about the distribution of the data. c) the estimates and the inferences for the estimates can be obtained simultaneously. For categorical data, Williamson et al. (2000) considered the case of measuring kappa between any two of the k raters, with each rater measuring each of the n subjects once. For continuous data, Barnhart et al. (2005) considered the case of measuring CCC, among any two and among all k raters, with each rater measuring each of the n subjects multiple times (independent replications). Barnhart et al. (2005) proposed a series of indices (intra-rater CCC, inter-rater CCC, and total CCC) and estimate those indices and their inferences by GEE method. For agreement among k raters, with each rater measuring each of the n subjects once, King and Chinchilli (2001a) proposed a generalized CCC, which can be reduced to kappa and weighted kappa for categorical data and original CCC for continuous data. For normal data where each rater measures each of the n subjects once, Carrasco and Jover (2003) recognized that the overall concordance correlation coefficient proposed by Lin (1989), Barnhart et al. (2002) and King and Chinchilli (2001a) are the same and it is a special version of the Intra-class Correlation (ICC) (Bartko, 1966; Fisher, 1925; Fleiss, 1986; Shrout and Fleiss, 1979). Carrasco and Jover (2003) proposed to estimate CCC by variance components method (Searle et al., 1992) with a two-way mixed no interaction model using maximum likelihood (ML) or restricted maximum likelihood (REML) approaches. Through their model, CCC can be used to measure the agreement among k raters, with each rater measuring each of the n subjects once. This paper proposes an approach which integrates the approaches by Barnhart et al. (2005) and Carrasco and Jover (2003).

This paper is structured as follows: in Section 2, we introduce a unified approach which can be used for continuous, binary, and ordinal data. In Section 3, we provide the simulation results in assessing the performance of the unified approach. In Section 4, we give two examples to illustrate the use of the unified approach. Finally, we draw conclusions and provide some discussions in Section 5.

2. METHOD

Suppose each of the k raters or methods measuring each of the n subjects m replicated times. The model we use for measuring agreement is

$$y_{ijl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijl} \tag{1}$$

Here, y_{ijl} stands for the lth reading from subject i given by rater j, with $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, and $l = 1, 2, \ldots, m$. The reading can be continuous, binary or ordinal response. μ is the overall mean. α_i is the random subject effect ($\sim(0, \sigma_\alpha^2)$) with equal second moments across raters. γ_{ij} is the random interaction effect between rater and subject ($\sim(0, \sigma_\gamma^2)$) with equal second moments across raters. Similarly, we assume e_{ijl} is the random error effect ($\sim(0, \sigma_e^2)$). β_j is the rater effect. Assume β_j is fixed and $\sum_{j=1}^{k} \beta_j = 0$. Even though β_j is a fixed effect, we still compute the variance among all raters, which is denoted as

$$\sigma_\beta^2 = \frac{\sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} (\beta_j - \beta_{j'})^2}{k(k-1)} \tag{2}$$

Based on the above model, we propose a series of indices to measure agreement, precision, and accuracy. We use $\bar{y}_{ij\cdot}$ to denote the average of m readings from subject i given by rater j, $\bar{y}_{i\cdot\cdot}$ to denote the average of all km readings from subject i, and $\bar{y}_{\cdot jl}$ to denote the average of n readings from rater j in its lth replication and y_{ijl} to denote any reading l from subject i given by rater j.

2.1. Intra-Rater Agreement

For a given rater, the intra-rater precision between any two replications, l and l' is

$$\begin{aligned} CCC_{intra} = \rho_{intra} &= \left. \frac{cov(y_{ijl}, y_{ijl'})}{\sqrt{(var(y_{ijl}))}\sqrt{(var(y_{ijl'}))}} \right|_{j(l,l')} \\ &= \frac{\sigma_\alpha^2 + \alpha_\gamma^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2}. \end{aligned} \tag{3}$$

Here, for each rater j, ρ_{intra} measures the proportion of the variance that is attributable to the subjects. Based on the proposed model, this proportion is the same for all k raters. Therefore, the intrarater agreement CCC_{intra} equals to $\rho_{intra.}$ This agreement index is heavily dependent on the total variability (total data range).

To examine the absolute agreement independent of the total data range, for $k = 2$ and $m = 1$, Lin (2000, 2003) and Lin et al. (2002) proposed two agreement indices based on mean squared deviation (MSD): Total Deviation Index (TDI) and Coverage Probability (CP), where TDI_π is

$$\delta_\pi = \phi^{-1}\left(1 - \frac{1-\pi}{2}\right)|\epsilon|, \tag{4}$$

and CP_δ is

$$\pi_\delta = 1 - 2[1 - \phi(\delta/|\epsilon|)] = x^2(\delta^2/\epsilon^2, 1), \tag{5}$$

ϵ^2 is MSD $= E(y_{il} - y_{i2})^2$ for $k = 2$, $m = 1$. ϕ is the cumulative normal distribution and $|.|$ is the absolute value. TDI (CP) is an approximate measure that captures a large proportion, CP, of data that are within a TDI boundary. Both TDI and CP depend on the normality assumption, and the approximations are good only when the relative bias square is small (Lin, 1992).

When measuring intra-rater agreement, we use ϵ^2_{intra} to denote the MSD between two replications l and l' for rater j.

$$\begin{aligned}\epsilon^2_{intra} &= E(y_{ijl} - y_{ijl'})^2 \\ &= (\mu_j - \mu_j)^2 + 2(\sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_e) - 2(\sigma^2_\alpha + \sigma^2_\gamma) \\ &= 2\sigma^2_e. \end{aligned} \tag{6}$$

Thus, $TDI_{intra(\pi)}$ can be expressed as

$$\delta_{intra(\pi)} = \phi^{-1}\left(1 - \frac{1-\pi}{2}\right)\sqrt{(2\sigma^2_e)}, \tag{7}$$

and $CP_{intra(\delta)}$ can be expressed as

$$\pi_{intra(\delta)} = 1 - 2\left[1 - \phi\left(\delta/\sqrt{(2\sigma^2_e)}\right)\right] \tag{8}$$

When the residual standard deviation becomes proportional to the measurement, we apply the natural log transformation of the data and then compute the agreement statistics. Under this situation, $TDI_\pi\%$ (anti-transform the TDI_π and subtract 1), which measures a per cent change rather than an absolute deviation, will be used. The $TDI_\pi\%$ is defined as

$$\delta_\pi\% = 100[\exp(\delta_\pi) - 1]\%. \tag{9}$$

To compute CP based on a percent change criterion, we need to convert the percent change to log scale based on Equation (9).

2.2. Inter-Rater Agreement

Since there are m replicated readings for subject i given by rater j, the average of those m readings could be used to measure the inter-rater agreement. Inter-rater agreement is a measure of agreement based on the average of multiple readings from each rater. Since readings from different raters have different expectations, inter-rater agreement CCC_{inter} consists of two parts: $precision_{inter}$ and $accuracy_{inter}$. The CCC becomes

$$\begin{aligned}\rho_{c,inter} &= 1 - \frac{E\left[\frac{\sum_{j=1}^k (\bar{y}_{ij.} - \bar{y}_{i..})^2}{(k-1)}\right]}{E\left[\frac{\sum_{j=1}^k (\bar{y}_{ij.} - \bar{y}_{i..})^2}{(k-1)} \mid \bar{y}_{i1.}, \bar{y}_{i2.}, \ldots, \bar{y}_{ik}, ind\right]} \\ &= \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \alpha^2_\gamma + \frac{\sigma^2_e}{m} + \sigma^2_\beta}. \end{aligned} \tag{10}$$

The precision index becomes

$$\rho_{inter} = \rho_{inter}|_{(j,j')} = \frac{cov(\bar{y}_{ij.}, \bar{y}_{ij'.})}{\sqrt{(var(\bar{y}_{ij.})}\sqrt{(var(\bar{y}_{ij'})}}\Big|_{jj'}$$
$$= \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \frac{\sigma_e^2}{m}}. \tag{11}$$

The accuracy index becomes

$$\chi_{a,inter} = \frac{\sigma_\alpha^2 + \sigma_\gamma^2 + \frac{\sigma_e^2}{m}}{\sigma_\alpha^2 + \sigma_\gamma^2 + \frac{\sigma_e^2}{m} + \sigma_\beta^2}. \tag{12}$$

the MSD becomes

$$\epsilon_{inter}^2 = 2\left(\sigma_\beta^2 + \sigma_\gamma^2 + \frac{\sigma_e^2}{m}\right). \tag{13}$$

The TDI and CP become

$$\delta_{inter(\pi)} = \phi^{-1}\left(1 - \frac{1-\pi}{2}\right)\sqrt{\left(2\sigma_\beta^2 + 2\sigma_\gamma^2 + 2\frac{\sigma_e^2}{m}\right)}, \tag{14}$$

and

$$\pi_{inter(\delta)} = 1 - 2\left[1 - \phi\left(\delta/\sqrt{\left(2\sigma_\beta^2 + 2\sigma_\gamma^2 + 2\frac{\sigma_e^2}{m}\right)}\right)\right]. \tag{15}$$

Here, $\rho_{c,inter}$ is the product of ρ_{inter} and $\chi_{a,inter}$. The accuracy index measures how close raters' means are. The definition of accuracy is some what different to that originally given by Lin (1989). Since in the proposed model, variances are assumed to be the same for different raters, the ratio of variances can't be included into the accuracy index. Note that the approach proposed by Barnhart et al. (2005) allows for different variances among raters. The inter-rater agreement is a measure of inter agreement based on the average of m readings made by each rater. Thus this index depends on the number of replications (m). The inter-CCC in Barnhart et al. (2005) is a measure of inter agreement based on the true readings from each rater. Thus it doesn't depend on the number of replications. That's why the inter-CCC from Barnhart et al. (2005) equals to the limit of our CCC_{inter} as the number of replications, m, goes to infinity. The approach proposed by Barnhart et al. (2005) allows for different intra rater coefficients.

2.3. Total Agreement

Since there are m replicated readings for subject i given by rater j, any one of the m replicated readings could be used to measure the inter-rater agreement. Total agreement is a measure of agreement based on any individual reading from

each reader. Thus this index does not depend on the number of replications. The corresponding indices are defined as:

$$\rho_{c,total} = 1 - \frac{E\left[\frac{\sum_{j=1}^{k}(y_{ijl}-\bar{y}_{i.l})^2}{(k-1)}\right]}{E\left[\frac{\sum_{j=1}^{k}(y_{ijl}-\bar{y}_{i.l})^2}{(k-1)} \mid y_{i1l}, y_{i2l}, \ldots, y_{ikl} ind\right]}$$
$$= \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2 + \sigma_\beta^2}, \tag{16}$$

$$\rho_{total} = \rho_{total}|_{(j,j')} = \frac{cov(y_{ijl}, y_{ij'l'})}{\sqrt{(var(y_{ijl}))}\sqrt{(var(y_{ij'l'}))}}$$
$$= \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2}, \tag{17}$$

$$\chi_{a,total} = \frac{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2 + \sigma_\beta^2}, \tag{18}$$

$$\epsilon_{total}^2 = 2\left(\sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2\right), \tag{19}$$

$$\delta_{total(\pi)} = \phi^{-1}\left(1 - \frac{1-\pi}{2}\right)\sqrt{\left(2\sigma_\beta^2 + 2\sigma_\gamma^2 + 2\sigma_e^2\right)}, \tag{20}$$

and

$$\pi_{total(\delta)} = 1 - 2\left[1 - \phi\left(\delta/\sqrt{\left(2\sigma_\beta^2 + 2\sigma_\gamma^2 + 2\sigma_e^2\right)}\right)\right] \tag{21}$$

2.4. Estimation and Inference

In order to estimate all indices and make related statistical inferences, the mean for each rater and all variance components, $\mu_1, \mu_2, \ldots, \mu_k$, σ_α^2, σ_e^2, σ_γ^2, and σ_β^2, need to be estimated first. Based on model (1) and balanced data, all variance components can be expressed as follows.

$$\sigma_e^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\sigma_{ij}^2}{nk}, \tag{22}$$

where σ_{ij}^2 stands for the conditional variance of Y_{ijl} given i and j.

$$\sigma_\alpha^2 = \frac{2\sum_{j=1}^{k-1}\sum_{j'=j+1}^{k}\sum_{l=1}^{m}\sum_{l'=1}^{m}\sigma_{jj'll'}}{m^2k(k-1)} \tag{23}$$

where $\sigma_{jj'll'}$ stands for the conditional covariance of Y_{ijl} and $Y_{ij'l'}$ given j, j', l and l'.

$$\sigma_\gamma^2 = A + B - C - D, \tag{24}$$

where

$$A = \frac{\sum_{j=1}^{k} \sum_{l=1}^{m} \sigma_{jl}^2}{m^2 k}, \tag{25}$$

and σ_{jl}^2 stands for the conditional variance of Y_{ijl} given j and l.

$$B = \frac{2 \sum_{j=1}^{k} \sum_{l=1}^{m-1} \sum_{l'=l+1}^{m} \sigma_{jll'}}{m^2 k}, \tag{26}$$

where $\sigma_{jll'}$ stands for the conditional covariance of Y_{ijl} and $Y_{iji'}$, given j, l, and l'.

$$C = \frac{2 \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} \sum_{l=1}^{m} \sum_{l'=1}^{m} \sigma_{jj'll'}}{m^2 k(k-1)} = \sigma_\alpha^2, \tag{27}$$

and

$$D = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} \sigma_{ij}^2}{mnk} = \frac{\sigma_e^2}{m}. \tag{28}$$

Above equations show that each of the four variance components can be expressed as functions of all variances and pairwised covariances. Thus even though in the proposed unified approach, we assume the homogeneity of all variances, the estimates are the same as the Overall Concordance Correlation Coefficient (OCCC) proposed by Lin (1989), King and Chinchilli (2001a), and Barnhart et al. (2002), where they didn't use the assumption of homogeneity of all variances.

The following system of equations (see Appendix A) are used to estimate each of rater means and all variance components, $\mu_1, \mu_2, \ldots, \mu_k, \sigma_\beta^2, \sigma_\alpha^2, \sigma_\gamma^2, \sigma_e^2$:

$$\sum_{i=1}^{n} F_i' H_i^{-1} (YY_i - \vartheta_i) = 0, \tag{29}$$

where

$$\mathbf{YY}_i = \begin{pmatrix} (y_{i11} + y_{i12} + \cdots + y_{i1m})/m \\ \cdots \\ (y_{ij1} + y_{ij2} + \cdots + y_{ijm})/m \\ \cdots \\ y_{ik1} + y_{ik2} + \cdots + y_{ikm}/m \\ \frac{1}{k(k-1)} \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} (\bar{y}_{ij.} - \bar{y}_{ij'.})^2 \\ \frac{2}{m^2 k(k-1)} \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} \sum_{l=1}^{m} \sum_{l'=1}^{m} [(y_{ijl} - \bar{y}_{.jl})(y_{ij'l'} - \bar{y}_{.j'l'})] \\ \frac{1}{k} \sum_{j=1}^{k} \left[\frac{\sum_{l=1}^{m} (y_{ijl} - \bar{y}_{ij.})^2}{(m-1)} \right] \\ \frac{1}{m^2 k} \sum_{j=1}^{k} \sum_{l=1}^{m} (y_{ijl} - \bar{y}_{.jl})^2 + \frac{2}{m^2 k} \sum_{j=1}^{k} \sum_{l=1}^{m-1} \sum_{l'=l+1}^{m} (y_{ijl} - \bar{y}_{.jl})(y_{iji'} - \bar{y}_{.jl'}) \end{pmatrix}$$

$$\vartheta_i = E(YY_i) = \begin{pmatrix} \mu_1 \\ \cdot \\ \mu_j \\ \cdot \\ \mu_k \\ \sigma_\beta^2 + \frac{\sigma_e^2}{m} + \sigma_\gamma^2 \\ \sigma_\alpha^2 \\ \sigma_e^2 \\ \sigma_\alpha^2 + \frac{\sigma_e^2}{m} + \sigma_\gamma^2 \end{pmatrix},$$

$H_i = diag(Var(YY_i))$, and $F_i = \frac{\partial \vartheta_i}{\partial(\mu_1,\ldots,\mu_k,\sigma_\beta^2,\sigma_\alpha^2,\sigma_\gamma^2,\sigma_e^2)}$. Based on the estimates and inferences estimates for all means and variances components, delta method is used to obtain the estimates and inferences of estimates for all indices (see Appendix B). When performing inferences on CCC-indices and precision indices, Z-transformation is used. When performing inferences on accuracy and CP indices, logit transformation is used. When performing inferences on TDIs, the natural log transformation is used.

For ordinal and binary data, when $k = 2$ and $m = 1$, the above GEE estimates of CCC reduce into kappa and weighted kappa with square distance function (Cohen, 1960, 1968; Robieson, 1999). In addition, its variances (Wu, 2005) reduce to the variances of kappa and weighted kappa (Fleiss et al., 1969).

3. SIMULATION STUDIES

In order to evaluate the performance of the GEE approach for estimation and inference of the proposed indices and to compare the proposed indices against other existing methods, simulation studies are conducted for different types of data: binary data, ordinal data, and normal data. For each of the three types of data, we consider three cases. Case one: k equals to 2 and m equals to 1. Case two: k equals to 4 and m equals to 1. Case three: k equals to 2 and m equals to 3. For each case, we generate 1000 random samples of size 20. For binary and ordinal data, we consider two situations: inferences obtained through transformations (z-transformations for CCC and precision indices, logit transformation for accuracy indices) and inferences obtained without transformations. For normal data, we only consider inferences obtained through transformations. In addition to the above transformation, we consider logit transformation for CP and log transformation for TDI. Simulation results are reported in Tables 1–5. For each table, the first column "THEORETICAL" stands for the theoretical value for this case. The second column "MEAN" stands for the mean of the 1000 estimated indices from the 1000 random samples. The comparisons between the first column and the second column are used to evaluate the robustness of the estimates. The third column "STD(EST)" stands for the standard deviation of the 1000 estimated indices from the 1000 random samples. The fourth column "MEAN(STD)" stands for the mean of the 1000 estimated standard errors. The comparison between the third and the fourth

Table 1 Binary data simulation results: with transformation

Stat	Theoretical	Mean	Std (Est)	Mean (Std)	Sig
Case one: $k = 2,\ m = 1$					
CCC	0.54992	0.56439	0.27613	0.25906	0.034
$precision$	0.59774	0.61211	0.26173	0.24256	0.072
$accuracy$	0.92000	0.92965	1.11350	1.14846	0.073
$CCC_{carrasco}$	0.54992	0.57826	0.28222	0.21477	0.04
Case two: $k = 4,\ m = 1$					
CCC	0.62739	0.62709	0.22486	0.20915	0.05
$precision$	0.66085	0.66820	0.21528	0.18338	0.051
$accuracy$	0.94937	0.94331	0.83210	1.02063	0.05
$CCC_{carrasco}$	0.62739	0.63851	0.22686	0.15799	0.051
Case three: $k = 2,\ m = 3$					
CCC_{inter}	0.78575	0.77510	0.13063	0.11326	0.045
CCC_{total}	0.68243	0.67135	0.13478	0.12264	0.048
$precision_{intra}$	0.79838	0.78680	0.09158	0.08958	0.039
$precision_{inter}$	0.80590	0.80040	0.11589	0.09804	0.047
$precision_{total}$	0.69758	0.68940	0.12561	0.11305	0.044
$accuracy_{inter}$	0.97500	0.96480	0.03698	0.03713	0.061
$accuracy_{total}$	0.97829	0.97023	0.03087	0.03820	0.052

column are used to evaluate the robustness of the variance estimates. The fifth column "SIG" stands for the proportion of estimates which are outside the 95% confidence interval at $\alpha = 0.05$.

For binary and ordinal data, we generate data from an underlying multivariate normal distribution and partition the responses into categories. When we generate the multivariate normal data, we specify the correlation in advance. When we partition the responses into categories, we specify the margins in advance. Thus, data sets with given precision and accuracy are generated. Those theoretical values are reported in each table with their theoretical values, denoted as "THEORETICAL".

For binary data with k equals to 2 and m equals to 1, we consider the correlation equals to 0.6. The margin for the first variable is (0.3, 0.7) and the margin for the second variable is (0.5, 0.5). For binary data with k equals to 4 and m equals to 1, we consider four variables x_1, x_2, x_3 and x_4 with vector mean $\mu = (0.55, 0.6, 0.65, 0.8)$ and $\rho_{12} = 0.75,\ \rho_{13} = 0.7,\ \rho_{14} = 0.5,\ \rho_{23} = 0.8,\ \rho_{24} = 0.6$, and $\rho_{34} = 0.6$. For binary data with k equals to 2 and m equals to 3, we consider six variables $x_{11}, x_{12}, x_{13}, x_{21}, x_{22}$, and x_{23} with vector mean $\mu = (0.7, 0.7, 0.7, 0.6, 0.6, 0.6)$. The correlation between any two of the first three variables is 0.8. The correlation between any two of the last three variables is also 0.8. The correlation between any one of the first three variables with any one of the last three variables is 0.7. The simulation results for binary data are reported in Tables 1 and 2. In Table 1, all estimates for standard deviations are values with transformations. All means are obtained through anti-transformations. For all cases in both Tables 1 and 2, our estimates are very close to their corresponding theoretical values, and the means of the estimated standard deviations are very close to their corresponding standard deviations of the estimates. Therefore, our estimates

Table 2 Binary data simulation results: without transformation

Stat	Theoretical	Mean	Std (Est)	Mean (Std)	Sig
Case one: $k = 2,\ m = 1$					
CCC^*	0.54992	0.53079	0.17595	0.16341	0.053
$precision$	0.59774	0.58072	0.15967	0.14399	0.057
$accuracy$	0.92000	0.89841	0.08087	0.07857	0.055
$CCC_{carrasco}$	0.54992	0.54021	0.17323	0.14042	0.054
Case two: $k = 4,\ m = 1$					
CCC	0.62739	0.61950	0.11996	0.12000	0.046
$precision$	0.66085	0.65969	0.10788	0.09753	0.045
$accuracy$	0.94937	0.93468	0.04101	0.05507	0.05
$CCC_{carrasco}$	0.62739	0.63008	0.11724	0.09293	0.046
Case three: $k = 2,\ m = 3$					
CCC_{inter}	0.78575	0.80198	0.34795	0.30357	0.047
CCC_{total}	0.68243	0.69016	0.25987	0.24204	0.052
$precision_{intra}$	0.79838	0.8059	0.27016	0.25605	0.054
$precision_{inter}$	0.8059	0.82369	0.33486	0.28528	0.047
$precision_{total}$	0.69758	0.70699	0.25189	0.23184	0.047
$accuracy_{inter}$	0.975	0.9796	1.3366	1.5165	0.092
$accuracy_{total}$	0.97829	0.9825	1.31543	1.48914	0.091

CCC^*: values are the same as the kappa, both in estimation and in inference.

Table 3 Ordinal data simulation results: with transformation

Stat	Theoretical	Mean	Std (Est)	Mean (Std)	Sig
Case one: $k = 2,\ m = 1$					
CCC	0.68569	0.68904	0.19127	0.17947	0.036
$precision$	0.78193	0.78494	0.15296	0.13975	0.036
$accuracy$	0.87692	0.88292	0.70825	0.65015	0.033
$CCC_{carrasco}$	0.68569	0.69977	0.19281	0.20136	0.036
Case two: $k = 4,\ m = 1$					
CCC	0.61551	0.61753	0.17087	0.16537	0.058
$precision$	0.65901	0.65821	0.17336	0.15873	0.059
$accuracy$	0.93398	0.93086	0.72901	0.83133	0.048
$CCC_{carrasco}$	0.61551	0.62925	0.17280	0.15681	0.058
Case three: $k = 2,\ m = 3$					
CCC_{intra}	0.85000	0.84673	0.19925	0.18690	0.045
CCC_{inter}	0.80573	0.81193	0.26196	0.23396	0.052
CCC_{total}	0.72756	0.72674	0.20410	0.18719	0.056
$precision_{intra}$	0.85000	0.84673	0.19925	0.18690	0.045
$precision_{inter}$	0.83333	0.84398	0.25327	0.22154	0.05
$precision_{total}$	0.75000	0.75273	0.19698	0.17875	0.06
$accuracy_{inter}$	0.96688	0.97306	1.34000	1.43990	0.072
$accuracy_{total}$	0.97009	0.97593	1.32621	1.42840	0.07

Table 4 Ordinal data simulation results: without transformation

Stat	Theoretical	Mean	Std (Est)	Mean (Std)	Sig
Case one: $k = 2$, $m = 1$					
CCC^*	0.68569	0.66761	0.10094	0.09282	0.042
precision	0.78193	0.77217	0.06496	0.05549	0.053
accuracy	0.87692	0.86205	0.07033	0.06783	0.047
$CCC_{carrasco}$	0.68569	0.67839	0.09942	0.10454	0.042
Case two: $k = 4$, $m = 1$					
CCC	0.61551	0.60432	0.10504	0.10036	0.05
precision	0.65901	0.64619	0.09943	0.08990	0.054
accuracy	0.93398	0.91750	0.04792	0.05170	0.039
$CCC_{carrasco}$	0.61551	0.61588	0.10394	0.09436	0.049
Case three: $k = 2$, $m = 3$					
CCC_{inter}	0.80573	0.79031	0.09275	0.08441	0.042
CCC_{total}	0.72756	0.71032	0.09685	0.09056	0.044
$precision_{intra}$	0.85000	0.83862	0.05687	0.05345	0.034
$precision_{inter}$	0.83333	0.82520	0.07890	0.06980	0.04
$precision_{total}$	0.75000	0.73789	0.08756	0.08060	0.043
$accuracy_{inter}$	0.96688	0.95614	0.03920	0.03710	0.057
$accuracy_{total}$	0.97009	0.96093	0.03453	0.03380	0.055

CCC^*: values are the same as the weighted kappa with mean squared weight, both in estimation and in inference.

are sufficiently good for binary data. We also report the estimates from Carrasco's method (Carrasco and Jover, 2003) for the same 1000 samples for cases one and two. The estimates are denoted as $CCC_{carrasco}$. For cases of $m = 1$ (cases one and two), our standard error estimates are superior to the estimates from Carrasco's method (Carrasco and Jover, 2003) regardless if we use transformation or not. We point out that the estimates from Table 1 with the estimates from Table 2, the estimates and their inference are comparable. Therefore, we suggest that for binary data, a transformation is not necessary.

For ordinal data with k equals to 2 and m equals to 1, we consider the correlation equals to 0.8, the margin for the first variable is (0.3, 0.3, 0.4) and the margin for the second variable is (0.25, 0.35, 0.4). For ordinal data with k equals to 4 and m equals to 1, we consider four variables x_1, x_2, x_3, and x_4 with margins: (0.3, 0.3, 0.4), (0.25, 0.35, 0.4), (0.2, 0.3, 0.5), and (0.4, 0.4, 0.2). The correlations among all variables are: $\rho_{12} = 0.7$, $\rho_{13} = 0.6$, $\rho_{14} = 0.75$, $\rho_{23} = 0.8$, $\rho_{24} = 0.6$, $\rho_{34} = 0.5$. For ordinal data with k equals to 2 and m equals to 3, we consider six variables $x_{11}, x_{12}, x_{13}, x_{21}, x_{22}$, and x_{23}. The first three variables had the same margin (0.3, 0.3, 0.4). The last three variables had the same margin (0.2, 0.3, 0.5). The correlation between any two of the first three variables is 0.85. The correlation between any two of the last three variables is also 0.85. The correlation between any one of the first three variables with any one of the last three variables is 0.75. The simulation results for ordinal data are reported in Tables 3 and 4, with Table 3 reporting the results with transformation and Table 4 without transformation. For both tables, the means of the estimates are very close to the theoretical values, and the means of the estimated standard errors are very close to the corresponding

Table 5 Normal data simulation results: with transformation

Stat	Theoretical	Mean	Std (Est)	Mean (Std)	Sig
Case one: $k = 2,\ m = 1$					
CCC	0.93101	0.9273	0.2267	0.19793	0.049
$precision$	0.95	0.95056	0.24215	0.204	0.047
$accuracy$	0.98001	0.98048	0.99008	1.28836	0.03
$TDI_{0.8}$	2.15056	2.08287	0.30864	0.27812	0.047
$CP_{2.15}$	0.79988	0.82004	0.48045	0.42595	0.048
$CCC_{carrasco}$	0.93101	0.93062	0.22699	0.21879	0.048
Case two: $k = 4,\ m = 1$					
CCC	0.855	0.84236	0.17386	0.16003	0.048
$precision$	0.9	0.89629	0.18422	0.16278	0.043
$accuracy$	0.95	0.94227	0.45674	0.4754	0.051
$TDI_{0.8}$	2.239	2.22074	0.17487	0.18929	0.046
$CP_{2.24}$	0.79998	0.80549	0.25189	0.21629	0.046
$CCC_{carrasco}$	0.855	0.84902	0.17468	0.16786	0.047
Case three: $k = 2,\ m = 3$					
CCC_{inter}	0.90663	0.9044	0.22513	0.1981	0.051
CCC^{***}_{inter}	0.92222	0.92286	0.26495	0.22882	0.05
CCC_{total}	0.87699	0.8711	0.19221	0.17166	0.045
$precision_{intra}$	0.95	0.94606	0.16939	0.15442	0.059
$precision_{inter}$	0.925	0.92649	0.23248	0.20222	0.048
$precision_{total}$	0.89417	0.89154	0.19165	0.17097	0.048
$accuracy_{inter}$	0.98014	0.98286	1.12096	1.40166	0.049
$accuracy_{total}$	0.98079	0.98351	1.11751	1.39846	0.049
$TDI_{intra(0.8)}$	1.98537	1.97069	0.15380	0.15054	0.056
$TDI_{inter(0.8)}$	2.69431	2.61924	0.32198	0.28825	0.045
$TDI_{total(0.8)}$	3.14437	3.09558	0.23245	0.21152	0.047
$CP_{intra(1.98)}$	0.79992	0.80607	0.22064	0.17155	0.059
$CP_{inter(2.69)}$	0.79995	0.81910	0.49259	0.33847	0.033
$CP_{total(3.14)}$	0.79995	0.81270	0.33810	0.31875	0.042

CCC^{***}_{inter}: calculated by Barnhart's method.

standard deviations of the estimates. Similar to binary data, we also report the estimates from Carrasco's method (Carrasco and Jover, 2003) for cases one and two in both tables. The estimates from two methods are very close to each other regardless of transformation being used or not. Thus we conclude that for ordinal data, transformation is not necessary. Carrasco's method (Carrasco and Jover, 2003) performs surprising well as ours for ordinal data.

For normal data with k equals to 2 and m equals to 1, we consider precision equals to 0.95, and accuracy equals to 0.98. For k equals to 4 and m equals to 1, we consider precision equals to 0.9 and accuracy equals to 0.95. For k equals to 2 and m equals to 3, we consider the within-rater precision equals to 0.95, between-rater precision equals to 0.925, and the between-rater accuracy equals to 0.98. The simulation results are reported in Table 5 with all standard errors obtained through transformed data. After we obtain the mean for the transformed data, we report its anti-transformation values in Table 5. Our estimates resemble their their theoretical values. Except for CP_{inter}, the means of the estimated standard error are very close to the corresponding standard deviations of the estimates. For cases of

$m = 1$, our CCCs are very close to that obtained from Carrasco's method (Carrasco and Jover, 2003). For case of $k = 2$ and $m = 3$, the inter-rater agreement calculated from Barnhart's (Barnhart et al., 2005) is a little bit larger than our inter-rater agreement since they assumed $m \to \infty$. Based on the proceeding simulations results, we conclude that our method works fairly well for binary data, ordinal data and normal data, both in estimates and in corresponding inferences.

4. EXAMPLES

In this section, we present two examples based on real data. Both examples consider replicated readings within each method or equipment.

4.1. Methods Comparison Example

Dispirin crosslinked hemoglobin (DCLHb) is a solution containing oxygen-carrying hemoglobin. The solution was created as a blood substitute to treat acute trauma patients and to replace blood loss during surgery. Measurements of DCLHb in patient's serum after infusion are routinely performed using a Sigma method. A method of measuring hemoglobin called the HemoCue photometer was modified to reproduce the Sigma instrument DCLHb results. To validate this modified method, serum samples from 299 patients over the analytical range of 50–2000 mg/dL were collected. DCLHb values of each sample were measured simultaneously with the HemoCue and Sigma methods and each sample were measured twice by each of the two methods. Similar method comparison examples have been given by Lin (2003) and Lin et al. (2002), where the averages of the replicated readings were used.

Figures 1 to 3 plot the data for this example: HemoCue method measure 1 vs. measure 2, Sigma method measure 1 vs. measure 2, the average of the HemoCue method vs. the average of the Sigma method. The plots indicate that the errors were rather constant across the data range. Therefore, no log transformation was applied to the data.

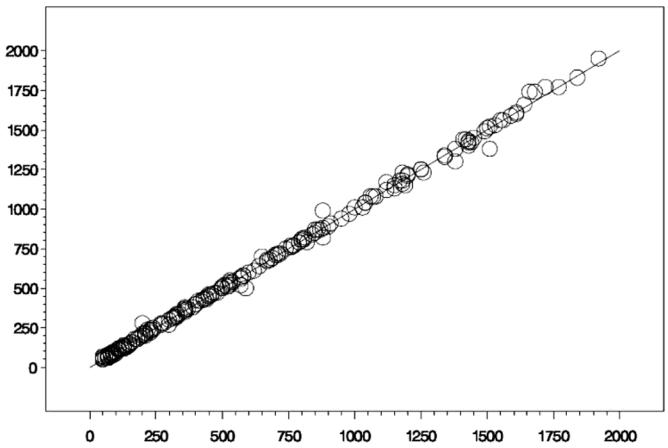

Figure 1 HemoCue method measure 1 vs. measure 2.

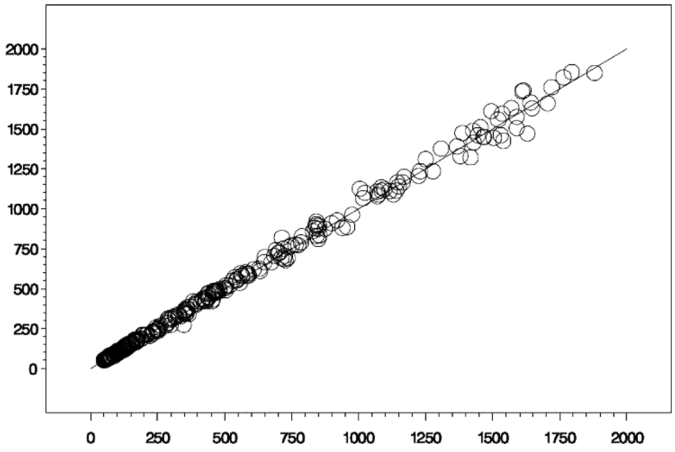

Figure 2 Sigma method measure 1 vs. measure 2.

In terms of TDI and CP indices, the least acceptable agreement is defined as having at least 90% of pair observations over the entire range within 75 mg/dL of each other if the observations are from the same method, and within 150 mg/dL of each other if the observations are from different methods based on the average of each method. In terms of CCC indices, the least acceptable agreement is defined as a within-sample total deviation not more than 7.5% of the total deviation if observations are from the same method, and a within-sample total deviation not more than 15% of the total deviation if observations are from different methods. These translates into a least acceptable CCC_{intra} of 0.9943 $(1 - 0.075^2)$, and a least acceptable CCC_{inter} of 0.9775 $(1 - 0.15^2)$.

The agreement statistics and their corresponding one-sided 97.5% lower or upper confidence limits for Example one are presented in Table 6. The CCC_{inter}

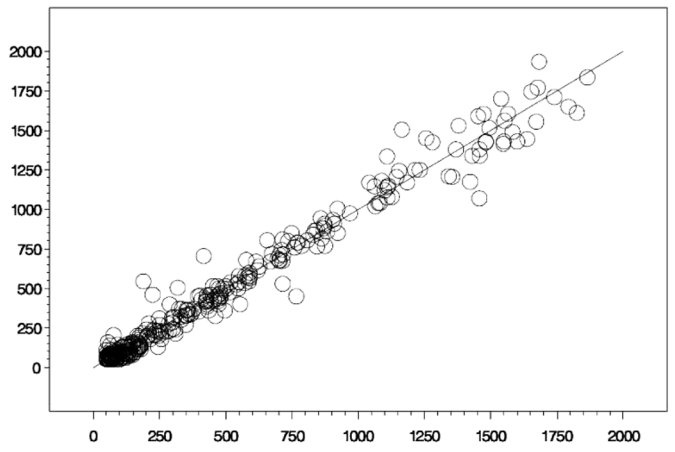

Figure 3 HemoCue method's average measure vs. Sigma method's average measure.

Table 6 Agreement statistics and their confidence limits for Example 1

Statistics	Estimates	97.5% Confidence limit*	Allowance
CCC_{inter}	0.9866	0.98153	0.9775
CCC_{total}	0.98592	0.98086	
$precision_{intra}$	0.99860	0.99823	0.9943
$precision_{inter}$	0.98664	0.98155	
$precision_{total}$	0.98595	0.98088	
$accuracy_{inter}$	0.99996	0.99742	
$accuracy_{total}$	0.99996	0.99742	
$TDI_{intra(0.9)}$	41.0903	47.2713	75
$TDI_{inter(0.9)}$	127.273	149.799	150
$TDI_{total(0.9}$	130.548	152.678	
$CP_{intra(75)}$	0.99732	0.99423	0.9
$CP_{inter(150)}$	0.94745	0.91701	0.9
$CP_{total(150)}$	0.94123	0.91016	

*: for all CCC, precision and accuracy indices, the 97.5% lower limits are reported, for all TDI indices, the 95% upper limits are reported.

estimate is 0.9866, which means for the average observations from different methods, the within-sample total deviation is about 11.6% ($\sqrt{(1-0.9866)}$) of the total deviations. The 97.5% lower confidence limit for CCC_{inter} is 0.9815, which is greater than 0.9775. The $precision_{intra}$ estimate is 0.9986, with a one-side lower confidence limit of 0.9982. The $precision_{inter}$ estimate is 0.98664 with a one-sided lower confidence limit of 0.9815, and the $accuracy_{inter}$ estimate is 0.99996 with one-sided lower confidence limit of 0.9974. The CCC_{total} estimate is 0.9859, which means for individual observations from different methods, the within-sample total deviation is about 11.87% of the total deviations. The 97.5% lower confidence limit for CCC_{total} is 0.9809. The $precision_{total}$ estimate is 0.9860 with a one-sided lower confidence limit of 0.9809, and the $accuracy_{total}$ estimate is 0.99996 with one-sided lower confidence limit of 0.9974. The $TDI_{intra(0.9)}$ estimate is 41.09 mg/dL, which means that 90% of the readings are within 41.09 mg/dL of their replicate readings from the same method. The one-sided upper confidence limit for $TDI_{intra(0.9)}$ is 47.2713, which is less than 75 mg/dL. The $TDI_{inter(0.9)}$ estimate is 126.16 mg/dL, which means based on the average readings, 90% of the readings are within 126.16 mg/dL of their replicate readings from the other method. The one-sided upper confidence limit for $TDI_{inter(0.9)}$ is 149.799 mg/dL, which is slightly less than 150 mg/dL. The $TDI_{total}(0.9)$ estimate is 130.55 mg/dL, with the one-sided upper confidence limit as 152.68 mg/dL. Finally, the $CP_{intra}(75)$ estimate is 0.9973, which means that 99% of HemoCue observations are within 75 mg/dL of their target values from same method. The one-sided lower confidence limit for $CP_{intra}(75)$ is 0.9942, which is larger than 0.9. The $CP_{inter}(150)$ estimate is 0.9475, which means that 95% of HemoCue observations are within 150 mg/dL of their target values from the other method based on the average of each method. The one-sided lower confidence limit for $CP_{inter}(150)$ is 0.9170, which is larger than 0.9. The $CP_{total}(150)$ estimate is 0.9412, which means that 94% of HemoCue observations are within 150 mg/dL of their target values from the other method based on individual readings. The one-sided lower confidence limit for $CP_{total}(150)$ is 0.9102.

Table 7 Lab one frequency table of first reading (row) vs. second reading (column)

	Negative	Positive	Highly positive
Negative	6	1	0
Positive	0	49	0
Highly positive	0	0	8

The agreement between HemoCue method and Sigma method is acceptable with acceptable precision and accuracy, with accuracy a little bit better than precision.

4.2. Assay Validation Example

In this example, we consider the Hemagglutinin Inhibition (HAI) assay for antibody to Influenza A (H3N2) in rabbit serum samples from two different labs. Serum samples from 64 rabbits are measured twice by each method. Antibody level is classified as: negative, positive, and highly positive (too numerous to count).

Tables 7 to 10 are the frequency tables for within lab and between lab readings. Tables 7 and 8 are frequency tables of the first reading vs. the second reading from each lab. Table 9 is the frequency table of the first reading from one lab vs. the first reading from the other lab. Table 10 is the frequency table of the second reading from one lab vs. the second reading from the other lab. Those tables suggest that the within lab agreement is good but the between lab agreement may not, and lab two tends to report higher values than lab one.

This is an imprecise assay, therefore we allow for looser agreement criteria. In terms of CCC indices, agreement was defined as a within-sample total deviation not more than 50% of the total deviation if observations are from the same method, and a within-sample total deviation not more than 75% of the total deviation if observations are from different methods. These translates into a least acceptable CCC_{intra} of 0.75 $(1 - 0.5^2)$, and a least acceptable CCC_{inter} of 0.4375 $(1 - 0.75^2)$.

The agreement statistics and their corresponding one-sided 97.5% lower confidence limits are presented in Table 11. The CCC_{intra} was estimated to be 0.88361, which means for observations from the same method, the within-sample deviation is about 34.1% $(\sqrt{(1 - 0.88361)})$ of the total deviations. The 97.5% lower confidence limit for CCC_{intra} is 0.79692, which is larger than 0.75. The CCC_{inter} is estimated to be 0.37225, which means for the average observations from different methods, the within-sample deviation is about 79.2% of the total

Table 8 Lab two frequency table of first reading (row) vs. second reading (column)

	Negative	Positive	Highly positive
Negative	2	0	0
Positive	0	22	2
Highly positive	0	5	33

Table 9 Lab one first reading (row) vs. lab two first reading (column)

	Negative	Positive	Highly positive
Negative	2	5	0
Positive	0	19	30
Highly positive	0	0	8

deviations. The 97.5% lower confidence limit for CCC_{inter} is 0.22039, which is less than 0.4375. The $precision_{inter}$ is estimated to be 0.56795 with a one-sided lower confidence limit of 0.4359, and the $accuracy_{inter}$ is estimated to be 0.65543 with a one-sided lower confidence limit of 0.51586. The CCC_{total} is estimated to be 0.35776, which means for individual observations from different methods, the within-sample deviation is about 80.1% of the total deviations. The 97.5% lower confidence limit for CCC_{total} is 0.2097. The $precision_{total}$ is estimated to be 0.53489 with a one-sided lower confidence limit of 0.3999, and the $accuracy_{total}$ was estimated to be 0.66885 with a one-sided lower confidence limit of 0.53561.

Overall, the agreement between two labs readings is not acceptable even though within lab agreement is much better than the inter lab agreement. The agreement within each lab can be obtained by applying kappa or weighted kappa to each lab separately.

5. DISCUSSION

We have proposed a series of indices for assessing agreement, precision and accuracy for the case of multiple raters each with multiple readings. Those indices can be used to assess intra, inter, and total agreement for both continuous and categorical data. All those indices are expressed as functions of variance components through a two-way mixed model, and GEE approach is used to estimate all indices and perform their inferences. All indices are summarized in Table 12.

All previously mentioned approaches for assessing agreement become one of the special case of our approach. For continuous data: when $m \to \infty$, the proposed estimates reduce to the agreement indices proposed by Barnhart et al. (2005). When $m = 1$, the proposed estimates reduce to the ICC proposed by Carrasco and Jover (2003). When $m = 1$, the proposed estimate also reduces to the OCCC proposed by Lin (1989), King and Chinchilli (2001a) and Barnhart et al. (2002). When $m = 1$ and $k = 2$, the proposed estimate reduces to the original CCC proposed by Lin (1989). For categorical data, when $k = 2$ and $m = 1$, the proposed estimate reduces

Table 10 Lab one second reading (row) vs. lab two second reading (column)

	Negative	Positive	Highly positive
Negative	2	4	0
Positive	0	23	27
Highly positive	0	0	8

Table 11 Agreement statistics and their confidence limits for Example 2

Statistics	Estimates	97.5% Confidence limit*	Allowance
CCC_{inter}	0.37225	0.22039	0.4375
CCC_{total}	0.35776	0.20970	
$precision_{intra}$	0.88361	0.79692	0.75
$precision_{inter}$	0.56795	0.43590	
$precision_{total}$	0.53489	0.39991	
$accuracy_{inter}$	0.65543	0.51586	
$accuracy_{total}$	0.66885	0.53561	

*: for all CCC, precision and accuracy indices, the 97.5% lower limits are reported.

to the kappa for binary data and weighted kappa with squared weight for ordinal data, in both estimates and inferences. In addition, we decompose the CCC into precision and accuracy components for a deeper understanding of the sources of the disagreement.

The concept of accuracy and precision can also be applied to categorical data. For continuous data, the above ICC like indices are heavily dependent on the total variability (total data range). Therefore, these indices are not comparable if the ranges of the data are not comparable. We also have proposed absolute indices, TDI and CP, which are independent of the total data range. These absolute indices are easily understandable by our clients. However, there absolute indices are valid only when the relative bias squared is small enough (Lin, 2000, 2003; Lin et al., 2002) and that the normality assumption is required.

Based on our unified approach, covariates adjustment can be easily applied by modifying the system of equations (27). The entire algorithm can be generalized to include and compare various functions of variance components.

Table 12 Summary of agreement indices based on functions of variance components

Statistics	Intra	Inter	Total	$m = 1$
CCC	$\frac{\sigma_\alpha^2+\sigma_\gamma^2}{\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2}$	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\gamma^2+\frac{\sigma_e^2}{m}+\sigma_\beta^2}$	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2+\sigma_\beta^2}$	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\beta^2+\sigma_e^2}$
Precision	$\frac{\sigma_\alpha^2+\sigma_\gamma^2}{\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2}$	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\gamma^2+\frac{\sigma_e^2}{m}}$	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2}$	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_e^2}$
Accuracy	NA	$\frac{\sigma_\alpha^2+\sigma_\gamma^2+\frac{\sigma_e^2}{m}}{\sigma_\alpha^2+\sigma_\gamma^2+\frac{\sigma_e^2}{m}+\sigma_\beta^2}$	$\frac{\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2}{\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2+\sigma_\beta^2}$	$\frac{\sigma_\alpha^2+\sigma_e^2}{\sigma_\alpha^2+\sigma_\beta^2+\sigma_e^2}$
MSD	$2\sigma_e^2$	$2\sigma_\beta^2+2\sigma_\gamma^2+2\frac{\sigma_e^2}{m}$	$2\sigma_\beta^2+2\sigma_\gamma^2+2\sigma_e^2$	$2\sigma_e^2+2\sigma_\beta^2$
TDI_π^*	$Q\sqrt{(MSD_{intra})}$	$Q\sqrt{(MSD_{inter})}$	$Q\sqrt{(MSD_{total})}$	$Q\sqrt{(MSD)}$
CP_δ^{**}	$\chi^2\left(\frac{\delta^2}{MSD_{intra}},1\right)$	$\chi^2\left(\frac{\delta^2}{MSD_{inter}},1\right)$	$\chi^2\left(\frac{\delta^2}{MSD_{total}},1\right)$	$\chi^2\left(\frac{\delta^2}{MSD},1\right)$

*: $Q=\phi^{-1}\left(1-\frac{1-\pi}{2}\right)$ is the inverse cumulative normal distribution **: $\chi^2\left(\frac{\delta^2}{MSD},1\right)$ is a central Chi-square distribution with one degree of freedom.

There are two aspects of this unified approach that can be developed in the future. First, for categorical and non-normal continuous data, we may include the link functions, such as, log or logit, in the GEE method. We expect the approach become more robust to different types of data after including link functions in the GEE method. Second, current variance components functions are based on balanced data, we can modify those functions to take care of the missing data.

For computing the above agreement statistics, a SAS macro is available (http://www.uic.edu/hedayat), which can be downloaded by users.

APPENDIX A: EXPRESSIONS FOR H_i AND F_i IN EQUATION (26)

For

$$\mathbf{YY}_i = \begin{pmatrix} (y_{i11} + y_{i12} + \cdots + y_{i1m}/m \\ \cdots \\ (y_{ij1} + y_{ij2} + \cdots + y_{ijm})/m \\ \cdots \\ y_{ik1} + y_{ik2} + \cdots + y_{ikm}/m \\ \frac{1}{k(k-1)} \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} (\bar{y}_{ij.} - \bar{y}_{ij'.})^2 \\ \frac{2}{m^2 k(k-1)} \sum_{j=1}^{k-1} \sum_{j'=j+1}^{k} \sum_{l=1}^{m} \sum_{l'=1}^{m} [(y_{ijl} - \bar{y}_{.jl})(y_{ij'l'} - \bar{y}_{.j'l'})] \\ \frac{1}{k} \sum_{j=1}^{k} \left[\frac{\sum_{l=1}^{m} (y_{ijl} - \bar{y}_{ij.})^2}{(m-1)} \right] \\ \frac{1}{m^2 k} \sum_{j=1}^{k} \sum_{l=1}^{m} (y_{ijl} - \bar{y}_{.jl})^2 + \frac{2}{m^2 k} \sum_{j=1}^{k} \sum_{l=1}^{m-1} \sum_{l'=l+1}^{m} (y_{ijl} - \bar{y}_{.jl})(y_{ijl'} - \bar{y}_{.jl'}) \end{pmatrix},$$

we have

$$H_i = diag(Var(YY_i)) = \begin{pmatrix} a & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & b & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & c & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & d & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & e & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & f & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & g \end{pmatrix},$$

where

$$a = var[(y_{i11} + y_{i12} + \cdots + y_{i1m})/m] = \sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2/m, \tag{30}$$

$$b = var[(y_{ij1} + y_{ij2} + \cdots + y_{ijm}/m] = \sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2/m, \tag{31}$$

$$c = var[(y_{ik1} + y_{ik2} + \cdots + y_{ikm}/m] = \sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2/m, \tag{32}$$

$$d = var\left[\frac{1}{k(k-1)}\sum_{j=1}^{k-1}\sum_{j'=j+1}^{k}(\bar{y}_{ij.}-\bar{y}_{ij'.})^2\right]$$
$$= \frac{k(k-1)(3k-2)}{m^2}\sigma_e^4 + k(k-1)(3k-2)\sigma_\gamma^4 + 8k(k-1)\sigma_\beta^2\sigma_\gamma^2$$
$$+ \frac{8k(k-1)}{m}\sigma_\beta^2\sigma_e^2 + \frac{4k^2(k-1)}{m}\sigma_\gamma^2\sigma_e^2, \tag{33}$$

$$e = var\left[\frac{2}{m^2k(k-1)}\sum_{j=1}^{k-1}\sum_{j'=j+1}^{k}\sum_{l=1}^{m}\sum_{l'=1}^{m}[(y_{ijl}-\bar{y}_{\cdot jl})(y_{ij'l'}-\bar{y}_{\cdot j'l'})]\right]$$
$$= m^4k(k-1)(2k-3)\sigma_\alpha^4 + \frac{m^4k(k-1)(2k-3)}{2}\sigma_\gamma^4 + \frac{m^2k(k-1)}{2}\sigma_e^4$$
$$+ \frac{[2+(m-1)^2+2m(m-1)(k-2)]k(k-1)m^2}{2}(\sigma_\alpha^2\sigma_e^2+\sigma_\gamma^2\sigma_e^2)$$
$$+ m^4k(k-l)(2k-3)\sigma_\alpha^2\sigma_\gamma^2, \tag{34}$$

$$f = var\left[\frac{1}{k}\sum_{j=1}^{k}\left[\frac{\sum_{l=1}^{m}(y_{ijl}-\bar{y}_{ij.})^2}{(m-1)}\right]\right]$$
$$= \frac{2k}{m-1}\sigma_e^4, \tag{35}$$

and

$$g = var\left[\frac{1}{m^2k}\sum_{j=1}^{k}\sum_{l=1}^{m}(y_{ijl}-\bar{y}_{\cdot jl})^2 + \frac{2}{m^2k}\sum_{j=1}^{k}\sum_{l=1}^{m-1}\sum_{l'=l+1}^{m}(y_{ijl}-\bar{y}_{\cdot jl})(y_{ijl'}-\bar{y}_{\cdot jl'})\right]$$
$$= [km(m-1)(2m-3) + m(m-1)(k-1)/2 + m^2k^2(3m-1)/2]\sigma_\alpha^4$$
$$+ [m^2k(1-3k/2+m+mk/2) + km(m-1)(2m-3)]\sigma_\gamma^4 + [mk(m-3)/2]\sigma_e^4$$
$$+ [2m^2k(2-k) + mk(m-1)(mk+5m-4)]\sigma_\alpha^2\sigma_\gamma^2$$
$$+ [mk(4-mk+(m-1)^2mk(m-1)/2]\sigma_\alpha^2\sigma_e^2$$
$$+ mk[4+(m-1)^2-mk+(m-1)(mk+4)/2]\sigma_\gamma^2\sigma_e^2. \tag{36}$$

we have

$$F_i = \frac{\partial\vartheta_i}{\partial(\mu_1,\ldots,\mu_k,\sigma_\beta^2,\sigma_\alpha^2,\sigma_e^2,\sigma_\gamma^2)}, \tag{37}$$

and

$$\mathbf{F}_i = \begin{pmatrix} 1_k & 0_{4*4} \\ 0_{4*4} & f_{4*4} \end{pmatrix},$$

where

$$f_{4*4} = \begin{pmatrix} 1 & 0 & 1/m & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1/m & 1 \end{pmatrix}.$$

APPENDIX B: VARIANCES FOR ALL INDICES

We use delta method to obtain the variances of estimates for all indices. These are:

$$\begin{aligned} &var(\rho_{c,\widehat{intra}}) \\ &\approx \frac{(1-\rho_{c,intra})^2[var(\sigma_\alpha^2)+var(\sigma_\gamma^2)+2cov(\sigma_\alpha^2,\sigma_\gamma^2)]+(\rho_{c,intra})^2 var(\sigma_e^2)}{(\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2)^2} \\ &\quad - \frac{2(1-\rho_{c,intra})(\rho_{c,intra})[cov(\sigma_\alpha^2,\sigma_e^2)+cov(\sigma_e^2,\sigma_\gamma^2)]}{(\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2)^2} \end{aligned} \tag{38}$$

$$\begin{aligned} &var(\rho_{c,\widehat{inter}}) \\ &\approx \frac{(1-\rho_{c,inter})^2 var(\sigma_\alpha^2)+(\rho_{c,inter})^2[var(\sigma_\beta^2)+var(\sigma_e^2)/(m^2)+var(\sigma_\gamma^2)}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2/m)^2} \\ &\quad + \frac{2cov(\sigma_\beta^2,\sigma_\gamma^2)+2cov(\sigma_\beta^2,\sigma_e^2)/m+2cov(\sigma_e^2,\sigma_\gamma^2)/m]}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2/m)^2} \\ &\quad - \frac{2(1-\rho_{c,inter})(\rho_{c,inter})[cov(\sigma_\beta^2,\sigma_\alpha^2)+cov(\sigma_\alpha^2,\sigma_\gamma^2)+cov(\sigma_\alpha^2,\sigma_e^2)/m]}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2/m)^2}, \end{aligned} \tag{39}$$

$$\begin{aligned} &var(\rho_{\widehat{inter}}) \\ &\approx \frac{(1-\rho_{inter})^2 var(\sigma_\alpha^2)+(\rho_{inter})^2[var(\sigma_e^2)/(m^2)+var(\sigma_\gamma^2)+2cov(\sigma_e^2,\sigma_\gamma^2)/m]}{(\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2/m)^2} \\ &\quad - \frac{2(1-\rho_{inter})(\rho_{inter})[cov(\sigma_\alpha^2,\sigma_\gamma^2)+cov(\sigma_\alpha^2,\sigma_e^2)/m]}{(\sigma_\alpha^2+\sigma_\gamma^2+\sigma_e^2/m)^2} \end{aligned} \tag{40}$$

$$\begin{aligned} &var(\rho_{c,\widehat{total}}) \\ &\approx \frac{(1-\rho_{c,total})^2 var(\sigma_\alpha^2)+(\rho_{c,total})^2[var(\sigma_\beta^2)+var(\sigma_e^2)+var(\sigma_\gamma^2)}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2)^2} \\ &\quad + \frac{2cov(\sigma_\beta^2,\sigma_\gamma^2)+2cov(\sigma_\beta^2,\sigma_e^2)+2cov(\sigma_e^2,\sigma_\gamma^2)]}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2)^2} \\ &- \frac{2(1-\rho_{c,total})(\rho_{c,total})[cov(\sigma_\beta^2,\sigma_\alpha^2)+cov(\sigma_\alpha^2,\sigma_\gamma^2)+cov(\sigma_\alpha^2,\sigma_e^2)]}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2)^2}, \end{aligned} \tag{41}$$

and

$$var(\rho_{\widehat{total}}) \approx \frac{(1-\rho_{total})^2 var(\sigma_\alpha^2) + (\rho_{total})^2[var(\sigma_e^2) + var(\sigma_\gamma^2) + 2cov(\sigma_e^2, \sigma_\gamma^2)]}{(\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2)^2}$$
$$\frac{-2(1-\rho_{total})(\rho_{total})[cov(\sigma_\alpha^2, \sigma_\gamma^2) + cov(\sigma_\alpha^2, \sigma_e^2)]}{(\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_e^2)^2} \tag{42}$$

When estimating the variances of CCC-indices and precision indices, we use the z-transformation, $Z_{index} = \frac{1}{2}\ln\frac{1+index}{1-index}$. Thus the transformed variances for all CCC and precision (ρ) indices are $var(z_{index}) = \frac{var(index)}{(1-(index)^2)^2}$, with the index being CCC_{intra}, CCC_{inter}, CCC_{total}, ρ_{inter}, or ρ_{total}.

When estimating the variances of TDI's, we use the log transformation based on MSD, $W = ln(\epsilon^2)$. The transformed variances for W become $\frac{var(\epsilon^2)}{\epsilon^4}$ For variance of MSD, we have

$$var(\epsilon^2_{\widehat{intra}}) = 4(var(\sigma_e^2)), \tag{43}$$

$$var(\epsilon^2_{\widehat{inter}}) = 4\left[var(\sigma_\beta^2) + var(\sigma_\gamma^2) + \frac{var(\sigma_e^2)}{(m^2)} + cov(\sigma_\beta^2, \sigma_\gamma^2) + \frac{cov(\sigma_\beta^2, \sigma_e^2)}{m} + \frac{cov(\sigma_e^2, \sigma_\gamma^2)}{m}\right] \tag{44}$$

and

$$var(\epsilon^2_{\widehat{total}}) = 4[var(\sigma_\beta^2) + var(\sigma_\gamma^2) + var(\sigma_e^2) + cov(\sigma_\beta^2, \sigma_\gamma^2) + cov(\sigma_\beta^2, \sigma_e^2) + cov(\sigma_e^2, \sigma_\gamma^2)]. \tag{45}$$

Note that TDI index is simply a scale transformation of the square root of MSD.

When estimating the variances of accuracy and CP indices, we use the logit transformation, $L_{index} = ln(\frac{index}{1-index})$ The transformed variances for accuracy or CP are $var(L_{index}) = \frac{var(index)}{(index)^2(1-index)^2}$ with index being $accuracy_{intra}$, $accuracy_{total}$, CP_{intra}, CP_{inter} or CP_{total}. For accuracy indices,

$$var(\chi_{a,\widehat{inter}}) = \frac{(1-\chi_{a,\widehat{inter}})^2[var(\sigma_\alpha^2) + var(\sigma_\gamma^2) + var(\sigma_e^2)/(m^2) + 2cov(\sigma_\alpha^2, \sigma_e^2)/m}{(\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2/m)^2}$$
$$+ \frac{2cov(\sigma_\alpha^2, \sigma_\gamma^2) + 2cov(\sigma_\gamma^2, \sigma_e^2)/m)] + (\chi_{a,inter})^2 var(\sigma_\beta^2)}{(\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2/m)^2}$$
$$- \frac{2(1-\chi_{a,inter})(\chi_{a,inter})[cov(\sigma_\alpha^2, \sigma_\beta^2) + cov(\sigma_\beta^2, \sigma_e^2)/m + cov(\sigma_\beta^2, \sigma_\gamma^2)]}{(\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2/m)^2} \tag{46}$$

and

$$var(\chi_{a,\widehat{total}}) = \frac{(1-\chi_{a,total})^2[var(\sigma_\alpha^2)+var(\sigma_\gamma^2)+var(\sigma_e^2)+2cov(\sigma_\alpha^2,\sigma_e^2)}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2)^2}$$
$$+\frac{2cov(\sigma_\alpha^2,\sigma_\gamma^2)+2cov(\sigma_\gamma^2,\sigma_e^2)]+(\chi_{a,total})^2var(\sigma_\beta^2)}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2)^2}$$
$$-\frac{2(1-\chi_{a,total})(\chi_{a,total})[cov(\sigma_\alpha^2,\sigma_\beta^2)+cov(\sigma_\beta^2,\sigma_e^2)+cov(\sigma_\beta^2,\sigma_\gamma^2)]}{(\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_e^2)^2} \tag{47}$$

For CP indices,

$$var(\hat{\pi}_{(.)}) = e^{\frac{\delta_{(\cdot)}^2}{\epsilon_{(\cdot)}^2}}\left(1+\frac{\delta_{(\cdot)}^2}{\epsilon_{(\cdot)}^2}\right)^2\frac{var(\epsilon_{(\cdot)}^2)}{8\pi\epsilon_{(\cdot)}^2\delta^2(\cdot)} \tag{48}$$

where $\epsilon_{(\cdot)}^2$, represents intra, inter and total MSD values shown in Equations (5), (13) and (18), respectively, and $var(\epsilon_{(\cdot)}^2)$ can be found in Equations (43)–(45) respectively.

We obtain the GEE estimates for above means and variances components as well as their variance–covariances interatively.

ACKNOWLEDGMENT

The research work for this article is supported by National Science Foundation (NSF) Grants DMS-0103727 and DMS-0603761, National Institutes of Health (NIH) Grant P50-AT00155 (jointly supported by National Center for Complementary and Alternative Medicine, the Office of Dietary Supplements, the Office of Research on Women's Health, and National Institute of General Medicine) and Astellas USA Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF and the NIH.

REFERENCES

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Pshchological Reports* 19:3–11.

Barnhart, H. X., Williamson, J. M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* 57:931–940.

Barnhart, H. X., Haber, M., Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 58:1020–1027.

Barnhart, H. X., Song, J., Haber, M. J. (2005). Assessing intra, inter and total agreement with replicated readings. *Statistics in Medicine* 19:255–270.

Carrasco, J. L., Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* 59:849–858.

Cohen, J. (1960). A coefficient of agreement for nomial scales. *Educational and Psychological Measurement* 20(1):37–46.

Cohen, J. (1968). Nomial scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4):213–220.

Fisher, R. A. (1925). *Statistical Methods for Researcher Workers*. Edinburgh: Oliver and Boyd.

Fleiss, J. L., Cohen, J., Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72:323–327.

Fleiss, J. L., Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33:613–619.

Fleiss, J. L. (1986). Reliability of measurement. *The Design and Analysis of Clinical Experiments*. New York: Wiley.

King, T. S., Chinchilli, V. M. (2001a). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* 20:2131–2147.

King, T. S., Chinchilli, V. M. (2001b). Robust estimators of the concordance correlation coefficient. *Journal of Biopharmaceutical Statistics* 11(3):83–105.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255–268.

Lin, L. I.-K. (1992). Assay validation using the concordance correlation coefficient. *Biometrics* 48:599–604.

Lin, L. I.-K. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 19:255–270.

Lin, L. I.-K., Hedayat, A. S., Sinha, B., Yang, M. (2002). Statistical methods in assessing agreement: models, issues and tools. *JASA* 97(457):257–270.

Lin, L. I.-K. (2003). Measuring agreement. *Encyclopedia of Biopharmaceutical Statistics* 561–567.

Robieson, W. Z. (1999). On Weighted Kappa and Concordance Correlation Coefficient. Ph.D. thesis, University of Illinois at Chicago.

Searle, R. S., Casella, G., McCulloch, C. E. (1992). *Variance Components*. New York: Wiley.

Shrout, P., Fleiss, J. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 86(2):420–428.

Williamson, J. M., Manatunga, A. K., Lipsitz, S. R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* 1(2):191–202.

Wu, W. (2005). A Unified Approach for Assessing Agreement. Ph.D. dissertation, University of Illinois at Chicago.

Copyright of Journal of Biopharmaceutical Statistics is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.