Statistical Methods in Medical Research 2006; **15**: 255–271

Coefficients of agreement for fixed observers

Michael Haber Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA, USA and **Huiman X Barnhart** Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, PO Box 17969, Durham, NC 27715, USA

Agreement between fixed observers or methods that produce readings on a continuous scale is usually evaluated via one of several intraclass correlation coefficients (ICCs). This article presents and discusses a few related issues that have not been raised before. ICCs are usually presented in the context of a two-way analysis of variance (ANOVA) model. We argue that the ANOVA model makes inadequate assumptions, such as the homogeneity of the error variances and of the pairwise correlation coefficients between observers. We then present the concept of observer relational agreement which has been used in the social sciences to derive the common ICCs without making the restrictive ANOVA assumptions. This concept did not receive much attention in the biomedical literature. When observer agreement is defined in terms of the difference of the readings of different observers on the same subject (absolute agreement), the corresponding relational agreement coefficient coincides with the concordance correlation coefficient (CCC), which is also an ICC. The CCC, which has gained popularity over the past 15 years, compares the mean squared difference between readings of observers on the same subject with the expected value of this quantity under the assumption of 'chance agreement', which is defined as independence between observers. We argue that the assumption of independence is unrealistic in this context and present a new coefficient that is not based on the concept of chance agreement.

1 Introduction

Commonly used coefficients for evaluation of agreement between observers making readings on a continuous variable are versions of the intraclass correlation coefficient (ICC). Over the past four decades, several review papers[1–4] presented summaries of the different types of ICCs and offered guidelines for the selection of the most appropriate ICC in a given situation. Despite this, there are still issues related to the definition and interpretation of coefficients of observer agreement that need clarification. In this article, we focus on two issues that, to our knowledge, have not been discussed in earlier works. We assume that the observers are fixed and that each observer makes at least one reading on each subject.

The first issue relates to the adequacy of the analysis of variance (ANOVA) model used to define coefficients of agreement between fixed observers. ICCs are usually defined in terms of variance components in a mixed two-way ANOVA model (random subjects, fixed observers). In Section 3, we will argue that this model makes assumptions that may not be appropriate when evaluating agreement between fixed observers, such as

Address for correspondence: Michael J Haber, Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA. E-mail: mhaber@sph.emory.edu

 10.1191/0962280206sm441oa

homogeneity of observer error variances and of interobserver correlations. Therefore, we present in Section 4 an alternative approach that, to our knowledge, has received very little attention in the biomedical literature. This approach is based on the concept of relational agreement which distinguishes different types of agreement. The most common type of agreement, at least in the biomedical sciences, is absolute agreement which means that the readings of different observers on the same subject should be identical, or at least close to each other. However, there are other types of relative agreement where the readings of one observer are only expected to follow a function of a specific type of those of another observer. For example, additive agreement between two observers means that each reading of the second observer can be obtained from the corresponding reading of the first by adding a constant. In Section 4, we will present a unified approach allowing different types of functional relationships and we will show that different types of agreement lead to different versions of the ICC. These coefficients of relational agreement are not based on the ANOVA model. Instead, they are defined in terms of the expected mean squared difference between readings made by pairs of observers on the same subject divided by its value under 'chance agreement'. Chance agreement is defined as independence (and hence lack of correlation) between observers. However, correlation and agreement are two different concepts. For example, consider two teachers who assign their students grades on a scale from 0 to 10 (Figure 1). In Figure 1(a), the grades are perfectly correlated while the absolute agreement is poor. In Figure 1(b), there is a good agreement while the correlation between the teachers

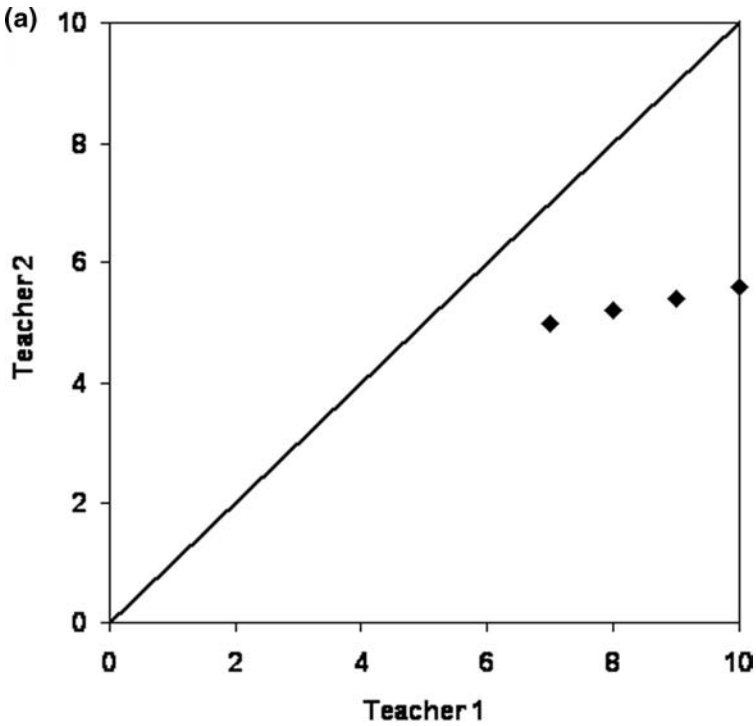

Figure 1 (a) Perfect correlation, poor agreement.

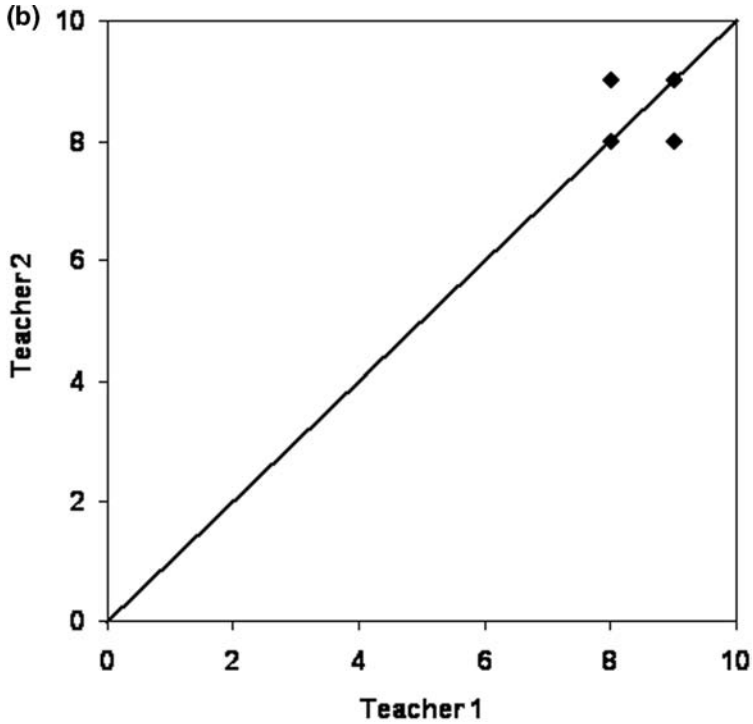

Figure 1 Continued. (b) Zero correlation, good agreement.

is zero. Therefore, the second issue discussed in this article (Section 5) relates to the appropriateness of this correction for chance agreement. We will argue that chance agreement should not be equated with independence and present in Section 6 a recently developed coefficient of interobserver agreement whose definition is not related to the concept of chance agreement.

Before we discuss the issues presented earlier, we provide in Section 2 a brief overview of the two commonly used coefficients of agreement for fixed observers. Both coefficients are versions of the ICC.

2 Definitions of ICCs for fixed observers

All the articles cited previously agree that the choice of an ICC depends on the type and origin of the data. Here, we will assume that each of J observers makes at least one observation on each of N subjects. We also assume that the observers are fixed, in the sense that we are only interested in agreement between these observers (rather than assuming that they represent a sample drawn from a large pool of observers). In this case, the two-way mixed-effect ANOVA model is commonly used to define the ICCs:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij} \tag{1}$$

This model presents a reading Y_{ij} of observer j on subject i as the sum of a random subject effect (α_i), a fixed observer effect (β_j), a random subject–observer interaction effect, (γ_{ij}) and a random error ε_{ij}. As we will see later, it is important to include a nonzero interaction effect, even though the variance of this effect cannot be estimated independently of the error variance unless observers make more than one reading on each subject. The sum of the interaction and error variances is always estimable. We denote the variances of the random effects associated with subjects, interaction and error by σ_S^2, σ_{SO}^2 and σ_E^2, respectively. In addition, we will denote the variability among the (fixed) observers' effects by $s_O^2 = \sum_j \beta_j^2/(J-1)$.

The classical definition of the ICC for this two-way model was given by Bartko[1]:

$$\mathrm{ICC_B} = \frac{\mathrm{Cov}(Y_{ij}, Y_{ij'})}{\sqrt{(\mathrm{Var}(Y_{ij}) \cdot \mathrm{Var}(Y_{ij'})}} = \frac{\sigma_S^2 - \sigma_{SO}^2/(J-1)}{\sigma_S^2 + \sigma_{SO}^2 + \sigma_E^2} \tag{2}$$

for $j' \neq j$. Actually, the expression given in Bartko[1] does not include the 'correction' $-\sigma_{SO}^2/(J-1)$ in the numerator of the right-hand side of Equation (2); however, in a later paper Bartko[5] pointed out that this correction should be included. This definition raises two questions that are not answered neither in Bartko's paper nor in most papers that present the identities in Equation (2) as a definition of the ICC:

- The first equality in Equation (2) states that $\mathrm{ICC_B}$ is identical to the ordinary product moment correlation coefficient (PMCC) between Y_{ij} and $Y_{ij'}$. What is the conceptual difference between Bartko's ICC and the PMCC? If we calculate $\mathrm{ICC_B}$ from the right-hand side of Equation (2), we usually obtain a number that is different from the PMCC. For example, if the readings of two observers on three subjects are (0,4), (5,5) and (10,6), then $\mathrm{ICC_B} = 0.38$ and PMCC $= 1.00$.[4]
- Suppose that there are at least three observers. The right-hand side of Equation (2) does not depend on which pair of observers is used, whereas the middle expression allows a different coefficient for each pair. How can these two expressions be equivalent?

The answer to both questions is the same: the two expressions in Equation (2) are equivalent only under the two-way model which assumes, among other things, that all the pairwise PMCCs are equal. Thus, the right-hand side of Equation (2) is the common correlation coefficient between two observers under the restrictions imposed by the two-way model. In the following section, we will explain why the two-way model may not be appropriate for the case of fixed observers.

From a practical point of view, the main question related to Bartko's definition [Equation (2)] is the following: what does $\mathrm{ICC_B}$ actually measure? As the right-hand side of Equation (2) does not include the variability between observers, it appears that this form of the ICC is not an appropriate measure of absolute agreement, that is, of the magnitude of the differences between the readings of different observers on the same study subject.

An answer to the last question can be found in McGraw and Wong.[4] They state that $\mathrm{ICC_B}$ attains its maximum value (one) if and only if the differences between the readings of any two observers are fixed across subjects, that is, for every pair of observers ($j \neq j'$),

$Y_{ij} - Y_{ij'}$ does not depend on i. However, if the ratio of the readings of two observers is fixed, then ICC_B may be considerably less than one. For example, for the readings (1,4), (2,8) and (3,12), $\text{ICC}_\text{B} = 0.47$. Thus, this form of the ICC is appropriate for measuring additive agreement (McGraw and Wong[4] labeled it as the consistency ICC).

To assess absolute agreement, McGraw and Wong suggested the correction of ICC_B by including the variability among the observers' main effects, s^2_O, in the denominator. Thus, they defined the agreement ICC as:

$$\text{ICC(agreement)} = \frac{\sigma^2_\text{S} - \sigma^2_\text{SO}/(J-1)}{\sigma^2_\text{S} + S^2_\text{O} + \sigma^2_\text{SO} + \sigma^2_\text{E}} \tag{3}$$

Another commonly used coefficient of absolute agreement is the concordance correlation coefficient (CCC), based on the chance-corrected expectation of the squared difference between the readings of two observers on the same subject. The CCC is usually attributed to Lin,[6] though this coefficient had been mentioned earlier by Krippendorff[7] and Zegers.[8] Barnhart *et al.*[9] extended the definition of the CCC to the case of multiple observers as follows:

$$\text{CCC} = 1 - \frac{E_i\{\sum_{j'>j}(Y_{ij} - Y_{ij'})^2\}}{E_i\{\sum_{j'>j}(Y_{ij} - Y_{ij'})^2 \text{ when } Y_1, Y_2, \ldots, Y_J \text{ are independent}\}} \tag{4}$$

where $\sum_{j'>j}$ stands for $\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}$. Barnhart *et al.*[9] showed that for the two-way model without interaction, the CCC is identical to the McGraw–Wong agreement ICC (3). Song[10] showed that the CCC is equivalent to the agreement ICC under the general two-way model with interaction. We will revisit the CCC in Sections 4 and 5, but we first examine the restriction imposed by the assumption that the observations follow a two-way ANOVA model.

3 Suitability of the two-way ANOVA model in assessing agreement between fixed observers

As we mentioned earlier, most approaches to assess observer agreement between fixed observers assume a two-way mixed ANOVA model (1) (the derivation of the CCC is not based on this model). Hence, it might be of interest to examine the suitability of the ANOVA model to the evaluation of agreement between fixed observers. Because we do not consider statistical inference in this article, we will not question the assumption of normality. Without any distributional assumptions, the two-way model requires that 1) all the observers have the same error variance and that 2) the correlation coefficients between all pairs of observers are the same. The first assumption implies that the error variance of all the observers is the same, which may not be true. The second assumption regarding equality of all pairwise correlations is even more questionable, as the correlation between two experienced observers is likely to be much higher than that between

an experienced and a novice observer. This is particularly disturbing because the ICCs are defined as types of correlation coefficients. It is also important to point out that the two-way ANOVA model without interaction assumes, among other things, that the differences between values assigned by two observers to each subject are fixed, apart from a random error. This assumption is unrealistic in many cases.

It should be pointed out that new multilevel linear models that allow estimation of variance components without the restrictive ANOVA assumptions are now available. These models can be fitted using the gllamm routine in the Stata software package.

Because of the aforementioned concerns regarding the appropriateness of two-way ANOVA models for observer agreement data, we looked for coefficients of agreement that are not based on these models. In the following section, we describe an alternative framework that can be used to derive coefficients of observer agreement. Two of these coefficients are equivalent to ICC_B and the CCC, respectively, whereas a third coefficient is the mean of all the pairwise PMCCs. A different and more general model for observer agreement data is presented in Section 6.

4 Observer relational agreement

The approach we are going to present for defining and measuring different types of observer agreement is based on the concept of observer relational agreement. This approach was introduced in the social sciences literature, but it has not received much attention in the biomedical/health sciences literature. The concept of relational agreement was introduced by Stine[11] for the case of two observers and generalized by Fagot[12,13] to any number of observers. It is based on the notion that prior to evaluating agreement, one has to define the appropriate scale of agreement. For example, if observers are expected to report the same value for each subject, then the corresponding scale is absolute agreement. If observers are allowed to differ from each other by a fixed value, but they are penalized for differences in their variances, then we are interested in agreement on the additive scale. Similarly, one may define agreement on the multiplicative, linear and ordinal scales. The most commonly used scale in the biomedical sciences is the absolute scale, as we are usually interested in determining the exact value of the quantity of interest for each subject.

In order to quantify this approach, we need to define a generic measure of agreement and a class of admissible transformations. A class of transformations is admissible with respect to a given scale of agreement when the measure of agreement attains its maximum value, that is, it indicates perfect agreement, if and only if the readings of one observer can be obtained from the corresponding readings of a second observer via a transformation from this class. For example, the measure of agreement on the additive scale should attain its maximum when the readings of one observer can be obtained from those of another by adding a constant. Therefore, the class of admissible transformations consists of all the additive transformations, $T(x) = x + a$. Similarly, the admissible transformations for absolute, multiplicative and linear agreements are $T(x) = x$, $T(x) = bx$ and $T(x) = a + bx$, respectively. As we want to quantify observer agreement, it is reasonable to limit ourselves to monotonically increasing transformations. Hence, for the multiplicative and

linear transformations, we require $b > 0$. The class of admissible transformations for ordinal agreement consists of all the monotonically increasing transformations.

It is important to point out that the scale of agreement does not have to be related to the scale of measurements. For example, suppose that the observed variable is the weight of an object in pounds, which is considered as measured on a ratio scale (as there is a unique well-defined zero). Then, the scale of agreement does not have to be multiplicative, as we usually want the actual measurements to be very close to each other, rather than allowing one measurement to be obtained from the other by multiplying by an arbitrary constant. In other words, we will be interested in absolute agreement, even though the scale of measurements is a ratio scale.

The second ingredient of the relational agreement approach is a function that determines the extent of agreement between observers. Usually it is required that the values of the function vary between -1 and 1, with 1 indicating perfect agreement, 0 indicating no agreement and -1 indicating 'perfect' negative agreement. The agreement function will be denoted by g. Consider an admissible transformation T (corresponding to a pre-determined scale of agreement) and denote $T_{ij} = T(Y_{ij})$. For the general case of J observers, Fagot[12] extended the earlier works by Zegers and ten Berge[14] and Stine[11] and proposed the following agreement function to determine the magnitude of agreement between $T_1, \ldots, T_J$, where $T_j = T(Y_j)$:

$$g(T_1, \ldots, T_J) = 1 - \frac{\sum_i \sum_{j'>j} (T_{ij} - T_{ij'})^2}{(J-1) \sum_i \sum_j T_{ij}^2} \tag{5}$$

Thus, $1 - g$ is the average Euclidean distance between the transformed readings of pairs of observers divided by a normalizing factor. Zegers and ten Berge[14] proposed the use of a specific transformation from each of the first four classes. These transformations, which are referred to as uniformed transformations, are listed in Table 1. In that table, $\bar{Y}_j$ and S_j^2 are the sample mean and variance, respectively, corresponding to the original readings of observer j. For the ordinal scale, Fagot[13] defined the uniformed transformation for Y_{ij} as the rank of that observation when the readings of observer j are ranked from smallest to largest. The various coefficients of relational agreement can be obtained by substituting T_{ij}s from Table 1 into Equation (5).

Table 1 Uniformed transformations for five scales of agreement

Scale of agreement	Uniformed transformation for Y_{ij}
Absolute	$T_{ij} = Y_{ij}$
Additive	$T_{ij} = Y_{ij} - \bar{Y}_j$
Multiplicative	$T_{ij} = Y_{ij} / \sqrt{\sum_i Y_{ij}^2 / N}$
Linear	$T_{ij} = (Y_{ij} - \bar{Y}_j)/S_j$
Ordinal	T_{ij} is the rank of subject i for observer j

Zegers[8] argued that the agreement function has to be corrected for 'agreement by chance'. He defined the chance-corrected agreement function as:

$$h = \frac{g - g_c}{1 - g_c} \tag{6}$$

where g_c is the expected value of g under chance agreement, which he interpreted as independence of the observers. It is calculated as the mean of g over $N!$ permutations of the readings of one observer, though the readings of the other observer remain fixed. Fagot[12] gave a general expression for the chance-corrected agreement function, $h(T_1, \ldots, T_J)$, which is obtained from Equations (5) and (6). For each of the first four scales of agreement in Table 1, Fagot gave an expression for h in terms of the sample moments of $Y_1, \ldots, Y_J$. The expressions for absolute, additive and linear chance-corrected agreement coefficients are as follows:

$$h_{AB} = \frac{2\sum_{j'>j} S_{jj'}}{(J-1)\sum_j S_j^2 + \sum_{j'>j}(\bar{Y}_j - \bar{Y}_{j'})^2} \tag{7}$$

$$h_{AD} = \frac{2\sum_{j'>j} S_{jj'}}{(J-1)\sum_j S_j^2} \tag{8}$$

$$h_L = \text{mean}_{j'>j}(r_{jj'}) \tag{9}$$

where $S_{jj'}$ and $r_{jj'}$ are the sample covariance and PMCC for the pair $(Y_j, Y_{j'})$. The expression for h_M, the chance-corrected coefficient of multiplicative agreement, is somewhat more complicated. For ordinal agreement, the corresponding coefficient is equivalent to the chance-corrected Kendal coefficient of concordance.[13]

We can substitute the population moments in the expressions for h by the corresponding sample moments and obtain population values (parameters) for the chance-corrected coefficients of relational agreement. We denote the mean and variance of Y_j by μ_j and σ_j^2, respectively. The PMCC of Y_j and $Y_{j'}$ is denoted by $\rho_{jj'}$ and the population covariances are $\sigma_{jj'} = \rho_{jj'}\sigma_j\sigma_{j'}$. The chance-corrected coefficients of absolute, additive and linear agreements are as follows:

$$\eta_{AB} = \frac{2\sum_{j'>j} \sigma_{jj'}}{(J-1)\sum_j \sigma_j^2 + \sum_{j'>j}(\mu_j - \mu_{j'})^2} \tag{10}$$

$$\eta_{AD} = \frac{2\sum_{j'>j} \sigma_{jj'}}{(J-1)\sum_j \sigma_j^2} \tag{11}$$

$$\eta_L = \text{mean}_{j'>j}(\rho_{jj'}) \tag{12}$$

The expression for the coefficient of multiplicative agreement is again somewhat more complicated.

How are these relational agreement coefficients related to the ICCs? First, the coefficient of absolute agreement equals to the CCC which, in turn, equals to the agreement ICC [Equation (3)] in terms of the parameters of the two-way model. Second, the coefficient of additive agreement equals to ICC_B.[12] Finally, the coefficient of linear agreement is the mean of all the pairwise PMCCs. Hence, we conclude that:

1) The two most commonly used ICCs in the case of fixed observers, namely Equations (2) and (3), can be derived from the concept of observer relational agreement and the agreement function (5), which is based on the distances between pairs of observers.
2) the PMCC is a valid coefficient of agreement when we define agreement as a linear association between observers.
3) these forms of the ICC can be derived and estimated without the assumption of a two-way ANOVA model.

4.1 Examples

To illustrate the three coefficients of relational agreement [Equations (7)–(9)], we use data from a study designed to determine the suitability of magnetic resonance angiography (MRA) for noninvasive screening of carotid artery stenosis, compared with invasive intra-arterial angiogram (IA). The main interest is in comparing two MRA techniques, two-dimensional (MRA-2D) and three-dimensional (MRA-3D) time of flight, to the IA, which is considered the 'gold standard'. In this example, the three screening methods are considered as the 'observers'. Readings were made by three raters using each of the three methods to assess carotid stenosis on each of 55 patients. Separate readings were made on the left and right carotid arteries. The data is presented in the appendix and can be copied from our website at: www.sph.emory.edu/observeragreement/data.htm.

In this section, we use the means of the readings made by the three raters on each subject using each method. The various coefficients of agreement are presented in Table 2. Obviously, the main interest in this example is in absolute agreement; the values of the additive and linear agreement coefficients are included for illustrative purposes only.

Table 2 Coefficients of relational agreement for the carotid stenosis data

Methods	Coefficient of agreement		
	Absolute	Additive	Linear
Left artery			
All three methods	0.668	0.683	0.683
IA, MRA-2D	0.675	0.685	0.685
IA, MRA-3D	0.556	0.582	0.582
MRA-2D MRA-3D	0.773	0.780	0.780
Right artery			
All three methods	0.743	0.772	0.773
IA, MRA-2D	0.762	0.815	0.816
IA, MRA-3D	0.689	0.723	0.724
MRA-2D MRA-3D	0.778	0.779	0.779

As expected, the absolute agreement coefficient is always the smallest, followed by additive and linear coefficients. As to the issue of which of the two MRA methods agrees better with the IA method, we see from Table 2 that for both the left and right carotid arteries, the absolute agreement coefficient of the MRA-2D and IA readings is greater than that of the MRA-3D and IA readings (0.675 versus 0.556 for the left artery and 0.762 versus 0.689 for the right artery).

Although in this example the three relational agreement coefficients do not differ substantially, other studies found more notable difference between the absolute and additive coefficients. For example, Muller and Buttner[3] present results from a study comparing two observers assessing cardiac output of 23 ventilated patients. For their data, the absolute agreement coefficient (CCC) is 0.751, the additive coefficient (Shrout and Fleiss's ICC for Model C) is 0.918 and the linear agreement coefficient (PMCC) is 0.924. Shrout and Fleiss[2] analyzed ratings of four judges on six subjects. The absolute, additive and linear agreement coefficients for their data are 0.284, 0.715 and 0.760, respectively.

Although it seems that these coefficients of relational agreement are appropriate for defining the extent of agreement between fixed observers, they suffer from a problem, which is discussed in the following section.

5 On the correction for chance agreement

The sample coefficients of relational agreement [Equations (7)–(9)] and the corresponding parameters [Equations (10)–(12)] are based on comparing the actual value of an agreement function with its expected value when the observers are independent (and therefore uncorrelated), because agreement by chance is interpreted as independence. As we stated earlier, lack of agreement does not imply, neither is it implied by lack of correlation. For the coefficients of additive and linear agreement, this does not constitute a problem, as the corrected and uncorrected coefficients are identical.[8] However, the coefficient of absolute agreement (CCC) is affected by the correction for chance agreement and therefore it may sometimes attain unexpected values. We will illustrate this using an example given by Zegers.[15] Suppose that two teachers assign grades (on a scale from 0 to 10) to each of four students and that the grades they assign to these students are (8,8), (8,9), (9,8), (9,9) [Figure 1(b)] Most of us would say that the agreement of these teachers is quite good. However, the CCC for this data is zero. This happens because the PMCC for these scores is zero and the absolute value of the CCC never exceeds that of the PMCC.[6] In other words, the correction for chance agreement used in the definition of the CCC leads to the misleading conclusion that lack of correlation always implied lack of agreement.

This unexpected behavior of the CCC results from the erroneous notion that agreement by chance means independence (which implies lack of correlation) between observers. Because the observers measure the same quantity on the same subject, it is highly unlikely that their readings will be independent. Of course, when each observer uses an independent random number generator to produce a 'reading' that is unrelated to the subject's true value, then the observers are independent, but this is not what we

perceive as agreement by chance. To us, agreement by chance means that observers are independent given the fact that they evaluate the same subject. In other words, we expect observers to agree simply because they observe the same subject. If this is the only source of dependence, then one may argue that the observers agree 'by chance'. If the level of agreement is beyond what is expected owing to the common value of the 'true' quantity, for example, because the observers use the same technique or because they have received the same training (or because one of them is not blind to the readings of another), then one can expect that their agreement is beyond chance agreement. In the following section, we present an alternative coefficient of absolute agreement that does not involve the dubious concept of chance agreement.

6 A new coefficient of interobserver agreement

From the previous sections, it is evident that a coefficient of absolute agreement between observers should be based on the mean squared difference between the readings of pairs of observers on the same subject. In other words, we agree that the numerator in expression (3) for the CCC, that is, $W^2 = E_i\{\sum_{j'>j}(Y_{ij} - Y_{ij'})^2\}$, is a valid measure of agreement, where a small value of W^2 indicates good agreement. The problem arises when we try to standardize W^2 so that we end up with a coefficient whose value is unity for 'maximum agreement' and zero for 'lack of agreement'. The standardization in the CCC is based on the value of W^2 in the case of lack of agreement. Because of the difficulties in defining lack of agreement, we decided to move to the opposite end of the agreement spectrum and use maximum agreement as a basis for standardization of W^2. How do we define maximum agreement? It seems that requiring $Y_{i1} = Y_{i2} = \cdots = Y_{iJ}$ for every i, that is, $W = 0$, would be too strict, as we should allow for observers' random errors. Therefore, we propose the following simple model:

$$Y_{ij} = \mu_{ij} + e_{ij}, \quad E(e_{ij}) = 0, \quad \mathrm{Var}(e_{ij}) = \sigma^2_{e,j} \tag{13}$$

Here, μ_{ij} can be regarded as the mean value (true value) that observer j would assign to subject i if the observer could make an infinite number of readings on this subject. The term e_{ij} is a random error, and the variance of this error, $\sigma^2_{e,j}$, represents the variability among replicated measurements that observer j makes (or would make) on this subject, that is, the intraobserver variability. We define maximum agreement as the minimum of W^2 over range of the true values μ_{ij} under model (13). Haber *et al.*[16] showed that the minimum is attained if and only if $\mu_{i1} = \mu_{i2} = \cdots = \mu_{iJ}$ for all i. Therefore, we propose the following coefficient of disagreement:

$$\delta = \frac{E_i\{\sum_{j'>j}(Y_{ij} - Y_{ij'})^2\}}{E_i\{\sum_{j'>j}(Y_{ij} - Y_{ij'})^2 \text{ when } \mu_{i1} = \cdots = \mu_{iJ} \text{ for every } i\}}$$

The reciprocal of this coefficient, $\psi = 1/\delta$, is our proposed new coefficient of absolute agreement. It varies between 0 and 1, with $\psi = 1$ indicating maximum agreement subject

to model (13). The other extreme case, $\psi = 0$, corresponds to a limiting situation where the variability among $\mu_{i1}, \ldots, \mu_{iJ}$ approaches infinity. In addition, $\psi = 0$ when there is no error, that is, $\sigma^2_{e,j} = 0$ for all j. This is reasonable because under $\sigma^2_{e,j} \equiv 0$, each observed value Y_{ij} equals to the corresponding true value μ_{ij}, and hence, any nonzero value of W^2 indicates disagreement for at least one pair of observers.

Haber *et al.*[16] also showed that ψ can be expressed as $1 - \tau^2/(\tau^2 + \sigma^2_e)$, where τ^2 is the mean (over subjects) of $\sum_j(\mu_{ij} - \bar{\mu}_i)^2$, and σ^2_e is the average (over observers) of the error variances $\sigma^2_{e,j}$. In other words, ψ is based on the ratio of the interobserver variability to the total observer variability. Therefore, ψ is a measure of interobserver agreement, whereas the coefficients discussed earlier measure the total observer agreement. The coefficient ψ can be estimated from the observed inter- and intraobserver variabilities.[16] Barnhart *et al.*[17] proposed CCC-like coefficients of interobserver and total observer agreements.

The main drawback of the approach presented in this section is the necessity to estimate the intraobserver variability. This variance component is best estimated when each observer makes two or more replicated measurements of each subject. When making replicated observations, it is important to make sure that the subjects' true values do not change between replications. This condition is satisfied when the subjects are x-ray slides, blood samples or verbal responses to an interview.

A similar approach has been used in studies on assessment of individual bioequivalence of two drugs. Schall and Luus[18] compare the difference in bioavailability $Y_T - Y_R$ between a test drug and a reference drug with the difference $Y_R - Y_{R'}$ between two readings on the reference drug. These differences correspond to the total and intra observer variabilities, respectively, in the context of the present article.

The coefficient ψ compares the observed value of W^2 to its expected values under the hypothesis $\mu_{i1} = \mu_{i2} = \cdots = \mu_{iJ}$ for all i. In other words, denoting by $\tilde{\mu}_j$ the random variable whose value on subject i is μ_{ij}, this hypothesis states that $P(\tilde{\mu}_j = \tilde{\mu}_{j'}) = 1$ for all $j < j'$. This hypothesis is stronger than the hypothesis of exchangeability, which states that the joint distribution of these J variables should be invariant under all the permutations of the indices $\{1, \ldots, J\}$.[19]

6.1 Examples

The carotid stenosis data, presented in Section 4 and in the appendix, consist of the readings of three raters with each of the three measurement methods. As we mentioned earlier, the three measurement methods are considered here as 'observers'. In addition, we now consider the readings by the three raters as replications. We calculated the ψ coefficients to assess the agreement among the three methods. The values of these coefficients for the left and right arteries are 0.632 and 0.738, respectively, not too far from the corresponding coefficients of absolute agreement (CCCs) of 0.668 and 0.743, respectively. In fact, one would expect the coefficient of interobserver agreement (ψ) to exceed the CCC, as the latter is a measure of the total observer agreement, which is based on both the inter- and intraobserver agreements. It is impossible to compare the two coefficients, in general, because the CCC depends on the total variability (interobserver, intraobserver

Table 3 Calcium scores on 12 patients

	Patient											
	1	2	3	4	5	6	7	8	9	10	11	12
A1	7	29	1	5	38	40	53	23	70	16	114	43
A2	6	31	1	6	32	29	49	23	70	15	116	43
B1	6	30	0	5	40	30	50	23	70	16	120	43
B2	6	30	0	5	40	29	51	24	70	16	120	43

and intersubject), whereas ψ depends on the inter- and intraobserver variabilities but not on the between-subject variability.

In order to better understand the effect of the between-subject variability on these measures of observer agreement, we consider the calcium scoring data presented by Haber *et al*.[16] Each of two radiologists, labeled A and B, made two replicated observations (labeled 1 and 2) on each of 12 patients. For this data, which is presented in Table 3, we obtained CCC $= 0.997$ (based on the mean of the two readings of each radiologist) and $\psi = 0.754$. In other words, the value of CCC indicates that there is perfect agreement between the two radiologists, whereas ψ indicates that there is notable disagreement. Examination of the data suggests that indeed there is some disagreement between the two radiologists (especially for patients 5, 6 and 11). The discrepancy between the two coefficients is mainly a result of the considerable between-subjects variability. This example demonstrates that the CCC may be unable to reflect observer disagreement when the between-subjects variability is substantially larger than the between-observer variability.[16]

The issue of the dependence of observer agreement coefficients on the between-subjects heterogeneity has been debated in the past. The CCC increases as the subjects are more heterogeneous, whereas ψ does not depend on this heterogeneity. For example, systolic blood pressure (SBP) has a larger variance than diastolic blood pressure (DBP); hence, the CCC for the former is usually higher than that for the latter. This would imply that it is more difficult to measure DBP than SBP, which is not supported by any real evidence.[3] For further discussions, see Atkinson and Nevill.[20]

7 Conclusions

The main conclusions from the above discussion are as follows:

- The concept of relational agreement provides a more appropriate framework for the derivation of coefficients of agreement between fixed observers when compared with the ANOVA approach, which imposes unrealistic and unnecessary restrictions.
- The correction for chance agreement used in the definition of the coefficient of absolute observer agreement (the CCC) is questionable, as chance agreement is not equivalent to independence between observers.
- The simple (and general) model (13) provides an alternative framework for derivation of coefficients of agreement. This model can be used when each observer makes two or more replicated readings on each subject.[16,17] Although the coefficient ψ

Table 4 Comparison of approaches used to define coefficients of agreement between fixed observers

Coefficients	Definition based on ANOVA assumptions?	Uses 'chance agreement' to scale the coefficient?	Depends on subjects' heterogeneity?	Requires replications?
ICCs	Yes[a]	No	Yes	No
Relational agreement including CCC	No	Yes	Yes	No
ψ	No	No	No	Yes

[a]The classical definition of the ICC is based on the variance components in the traditional ANOVA model. More recent multilevel linear models allow estimation of variance components when the ANOVA assumptions are relaxed.

is presented as a coefficient of absolute interobserver agreement, it can be used to measure other types of relational agreement by replacing the actual observations Y_{ij} by their transformed values T_{ij}.

Table 4 compares the three approaches that have been used to define the agreement coefficients discussed in this article.

The topic of assessing observer agreement is closely related to evaluation of measurement errors. A recent book by Dunn[21] summarizes various methods and approaches used in the analysis of data with measurement errors. Another recent book by Shoukri[22] reviews measures of observer agreement, though most of the book is devoted to categorical data.

Acknowledgement

This research was supported by NIMH grant 1 R01 MH070028-01 A1. The authors wish to thank two reviewers for their helpful comments.

References

1 Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.* 1966; **19**: 3–11.
2 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 1979; **86**: 420–28.
3 Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. *Stat. Med.* 1994; **13**: 2465–76.
4 McGraw KO, Wong SP. Forming inferences on intraclass correlation coefficients. *Psychol. Methods* 1996; **1**: 30–46.
5 Bartko JJ. Corrective note to: the intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.* 1974; **34**: 418.
6 Lin L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**: 255–68.
7 Krippendorff K. Bivariate agreement coefficients for reliability of data. *Sociol. Method.* 1970; **2**: 139–50.
8 Zegers FE. A family of chance-corrected association coefficients for metric scales. *Psychometrika* 1986; **51**: 559–62.
9 Barnhart HX, Haber M, Song J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* 2002; **58**: 1020–27.
10 Song J. *Assessing agreement/association for continuous measurement scales*. PhD thesis, Emory University, 2003.

11 Stine WW. Interobserver relational agreement. *Psychol. Bull.* 1989; **106**: 341–47.
12 Fagot RF. A generalized family of coefficients of relational agreement for numerical scales. *Psychometrika* 1993; **58**: 357–70.
13 Fagot RF. An ordinal coefficient of relational agreement for multiple judges. *Psychometrika* 1994; **59**: 241–51.
14 Zegers FE, ten Berge JMF. A family of association coefficients for metric scales. *Psychometrika* 1985; **50**: 17–24.
15 Zegers FE. Coefficients for interrater agreement. *Appl. Psychol. Meas.* 1991; **15**: 321–33.
16 Haber M, Barnhart HX, Song J, Gruden J. Observer variability: a new approach in evaluating interobserver agreement. *J. Data Sci.* 2005; **3**: 69–83.
17 Barnhart HX, Song J, Haber M. Assessing agreement in studies designed with replicated measurements. *Stat. Med.* 2005; **24**: 1371–84.
18 Schall R, Luus HG. On population and individual bioequivalence. *Stat. Med.* 1993; **12**: 1109–24.
19 Kelderman H. Measurement exchangeability and normal one-factor models. *Biometrika* 2004; **91**: 738–42.
20 Atkinson G, Nevill A. Comment on the use of concordance correlation to assess the agreement between variables. *Biometrics* 1997; **53**: 775–78.
21 Dunn G. *Statistical evaluation of measurement errors*, second edition. London: Arnold, 2004.
22 Shoukri MM. *Measures of interobserver agreement*. Boca Raton: Chapman & Hall/CRC, 2004.

Appendix

Appendix: Carotid artery stenosis data

Tables A1 and A2 present the carotid stenosis data for the left and right arteries of 55 patients. The data can be downloaded from our website: www.sph.emory.edu/observeragreement/data.htm.

Table A1 Carotid artery stenosis data–left artery

ID	IA			MRA-2D			MRA-3D		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
1	100	100	100	100	100	100	100	100	100
2	73	67	77	69	60	72	59	62	62
3	0	0	3	49	0	0	0	12	0
4	17	0	0	0	0	0	0	0	0
5	27	40	0	15	31	2	38	29	19
6	31	29	34	45	51	3	38	42	41
7	18	15	17	21	23	99	57	44	0
8	68	59	100	39	56	33	61	67	77
9	0	0	0	0	0	0	47	42	99
10	18	0	0	37	0	12	25	0	13

(*continued*)

Table A1 Continued

ID	IA			MRA-2D			MRA-3D		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
11	0	0	0	0	0	0	0	0	0
12	72	65	69	100	99	100	61	60	58
13	27	27	28	42	18	99	0	0	99
14	61	53	51	100	100	71	45	44	53
15	100	100	100	100	100	100	100	100	100
16	0	0	0	0	0	0	16	99	15
17	30	31	22	25	49	9	51	40	44
18	10	27	37	21	38	12	40	28	26
19	29	33	37	50	30	25	37	45	44
20	57	60	67	100	100	100	65	63	65
21	100	100	100	0	100	0	13	17	100
22	28	17	0	22	15	0	49	0	0
23	0	34	0	21	32	0	0	18	0
24	30	34	41	55	53	99	60	46	49
25	46	40	62	56	36	0	63	70	64
26	66	61	71	78	76	55	82	83	100
27	51	43	68	100	28	0	70	0	0
28	23	27	39	45	60	26	24	44	0
29	0	0	0	74	100	100	100	100	100
30	83	100	100	70	99	100	53	100	100
31	0	0	0	26	0	0	12	0	0
32	4	0	0	25	0	0	24	34	23
33	100	100	100	100	100	100	100	100	100
34	60	60	75	45	47	15	62	99	67
35	21	19	0	0	0	0	4	0	0
36	70	79	100	23	36	0	51	42	0
37	6	18	0	7	4	99	99	99	99
38	41	52	49	44	63	0	80	73	0
39	56	45	66	100	53	0	13	23	0
40	5	0	0	30	45	31	70	34	55
41	53	39	54	63	50	31	75	72	55
42	47	56	51	7	99	0	56	58	0
43	75	84	85	100	100	100	100	100	100
44	100	100	100	100	100	100	100	100	100
45	0	100	100	44	0	99	17	100	0
46	58	61	55	46	58	49	71	99	99
47	0	0	0	30	22	0	1	24	99
48	0	0	0	0	0	0	0	0	0
49	5	0	0	0	39	0	0	0	0
50	12	25	9	26	54	0	100	0	100
51	0	0	8	0	0	0	0	0	0
52	28	29	0	100	100	99	100	84	100
53	0	0	0	0	0	0	0	0	0
54	19	31	51	53	100	100	77	83	100
55	33	74	50	69	72	100	79	100	100

Table A2 Carotid artery stenosis data–right artery

ID	IA			MRA-2D			MRA-3D		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
1	32	26	23	36	49	32	36	26	17
2	28	30	23	43	0	2	36	18	41
3	7	17	8	34	0	99	14	22	44
4	0	0	0	24	0	0	0	28	0
5	22	23	0	27	99	0	11	17	99
6	78	100	100	100	100	100	33	100	100
7	67	63	78	100	100	38	56	20	0
8	64	74	76	100	100	100	61	41	40
9	12	0	0	39	62	33	100	100	100
10	100	100	100	100	100	100	100	100	100
11	59	53	31	14	26	12	29	30	26
12	11	3	13	0	99	0	0	0	0
13	100	100	100	100	100	100	100	100	100
14	76	69	70	100	100	100	77	81	75
15	40	61	100	70	100	100	100	100	100
16	4	0	0	9	0	0	18	0	0
17	39	53	52	57	51	52	74	64	75
18	6	32	0	0	36	1	0	13	7
19	0	0	0	31	22	2	49	55	54
20	5	11	9	7	22	0	0	7	0
21	4	11	0	18	16	0	99	11	99
22	17	35	33	15	44	0	28	0	0
23	0	0	0	0	0	0	0	0	0
24	25	51	28	37	63	34	100	41	25
25	55	60	65	65	41	7	68	100	6
26	13	20	0	26	0	0	11	21	0
27	44	36	58	44	35	14	68	28	33
28	50	62	53	52	62	67	57	51	51
29	44	24	42	56	45	45	52	58	58
30	51	45	47	49	50	47	36	100	56
31	4	28	0	14	0	49	0	0	0
32	0	22	7	2	31	0	41	40	0
33	15	0	0	17	99	99	24	40	0
34	0	49	0	42	47	100	15	38	99
35	80	84	83	100	100	100	100	100	100
36	40	22	41	51	33	17	11	0	0
37	43	45	44	41	99	99	41	99	99
38	23	0	20	1	23	0	9	20	0
39	1	0	0	40	99	0	0	99	99
40	14	17	20	36	27	0	0	50	14
41	26	31	34	46	51	99	23	100	21
42	0	0	0	0	0	0	13	0	0
43	100	100	100	100	100	100	100	100	100
44	0	0	0	26	0	0	0	0	0
45	0	0	0	45	0	99	27	99	0
46	57	52	65	39	62	0	62	62	28
47	0	20	0	40	99	5	20	21	7
48	0	0	0	6	40	0	0	0	0
49	11	0	0	23	55	0	3	36	0
50	32	27	30	10	34	20	8	49	99
51	100	100	100	100	100	100	100	100	100
52	62	63	71	80	81	100	64	77	68
53	12	19	13	43	16	31	21	39	27
54	63	67	84	100	100	100	72	99	99
55	27	39	46	63	63	67	65	65	73