Journal of Data Science **3**(2005), 69-83

Observer Variability: A New Approach in Evaluating Interobserver Agreement

Michael Haber[1], Huiman X. Barnhart[2], Jingli Song[3] and James Gruden[1]
[1]*Emory University,* [2]*Duke University and* [3]*Eli Lilly and Company*

Abstract: Existing indices of observer agreement for continuous data, such as the intraclass correlation coefficient or the concordance correlation coefficient, measure the *total* observer-related variability, which includes the variabilities between and within observers. This work introduces a new index that measures the *interobserver* variability, which is defined in terms of the distances among the 'true values' assigned by different observers on the same subject. The new coefficient of interobserver variability (CIV) is defined as the ratio of the interobserver and the total observer variability. We show how to estimate the CIV and how to use bootstrap and ANOVA-based methods for inference. We also develop a coefficient of excess observer variability, which compares the total observer variability to the expected total observer variability when there are no differences among the observers. This coefficient is a simple function of the CIV. In addition, we show how the value of the CIV, estimated from an agreement study, can be used in the design of measurements studies. We illustrate the new concepts and methods by two examples, where (1) two radiologists used calcium scores to evaluate the severity of coronary artery arteriosclerosis, and (2) two methods were used to measure knee joint angle.

Key words: Agreement studies, ANOVA, measurement studies, observer agreement, observer variability, random effects models.

1. Introduction

In a typical observer agreement study one is interested in comparing the readings made by the same J observer on each of I subjects. When the variable of interest is continuous, observer agreement is usually evaluated via one of the many versions of the intraclass correlation coefficient (Bartko, 1966; McGraw and Wong, 1996) or the concordance correlation coefficient (Lin, 1989). We believe that the choice of a proper measure of observer agreement should be based upon the investigator's objectives. In this work we assume that (i) the ultimate goal is to obtain an accurate and precise estimate of the true value of the variable of interest for *each* study subject (as opposed to, for example, estimating the

population mean), and (ii) the investigator is interested in assessing the actual differences *between* observers. In Section 2 we show that the existing agreement coefficients measure the *total* observer-related variability (or disagreement), which is the sum of the inter- and intra-observer sources of variability. Since we are mainly interested in the *between* observer disagreements, we introduce a new coefficient based on the *interobserver variability* which is defined as the mean (over subjects) sum of the squared distances of the 'true' measurements made by different observers on the same subject from the mean (over observers) of these 'true' measurements. The 'true' measurement of an observer on a subject is defined as the mean of all the values that would be assigned by the observer to the subject if the observer could make an infinite number of replicated measurements on that subject.

Evaluation of the *interobserver* variability is important when one is interested in the 'true' differences among observers reporting different values of the same quantity. In other words, the interobserver variability, rather than the total observer variability, should be used to explore the causes of disagreements among observers. The *total* observer variability masks these sources of disagreement as it contains both *interobserver* variability (true differences among observers) and *intraobserver* variability (random error among the observations made by the same observer on the same subject).

We define the *coefficient of interobserver variability* (CIV) as the ratio of the interobserver variability to the total observer-related variability. It varies between 0 and 1, and a higher value of the CIV indicates a lower level of interobserver agreement. If $CIV = 0$, then one does not expect any 'true' differences among the observers, in the sense that all the observers have the same distribution over the subjects. The quantity 1-CIV can be used as a coefficient of interobserver agreement.

Section 2 highlights the differences between the proposed coefficient and existing coefficients of agreement for continuous data. The definition and estimation of the CIV are presented in Section 3. We provide different (but very similar) coefficients for the fixed and random observers situations, and we show that in both cases the CIV is estimated by the same statistic. In Section 4 we introduce the coefficient of excess observer variability, defined as that ratio of the actual total observer variability to the expected value of this variability under the assumption of no true differences among observers. This coefficient is a simple function of the CIV. Section 5 shows how the CIV can be applied to help in designing studies that use measurements made by different observers in order to estimate the subjects' true values. In Section 6 we provide further understanding of the proposed concepts and methods when observations follow a two-way ANOVA model. These concepts and methods are illustrated in Section 7 by two

examples where (1) two radiologists used calcium scores to evaluate the severity of coronary artery arteriosclerosis, and (2) two methods were used to measure knee joint angle. We conclude with a discussion in Section 8.

2. The Need for a New Coefficient of Agreement

Traditionally, observer agreement has been measured via the intraclass correlation coefficient (ICC). The ICC was introduced by Fisher (1925) as a measure of the correlation between measurements made on pairs of brothers. Bartko (1966) introduced three versions of the ICC for the evaluation of observer reliability. However these, as well as most other forms of the ICC that have been proposed since then, measure correlation between observers rather than differences between observations made by different observers. McGraw and Wong (1996) presented an excellent summary of the various versions of the ICC and pointed out that one version, which we will refer to as the 'agreement ICC', is appropriate for evaluating agreement among observers. Using the ANOVA terminology, we denote the variabilities attributed to subjects, observers, subject by observer interactions and error (within observer) by $\sigma_S^2, \sigma_0^2, \sigma_{S0}^2$ and σ_E^2 respectively. Then the between and within observer variabilities are $\sigma_0^2 + \sigma_{S0}^2$ and σ_E^2, respectively, and the total observer-related variability is $\sigma_0^2 + \sigma_{S0}^2 + \sigma_E^2$. The agreement ICC is defined as $\sigma_S^2/(\sigma_S^2 + \sigma_0^2 + \sigma_{S0}^2 + \sigma_E^2)$. Hence, the agreement ICC compares the total observer variability to the between subjects variability and therefore it measures the total observer agreement. Lin (1989) introduced the concordance correlation coefficient (CCC) between two observers. Barnhart *et al.* (2002) generalized the CCC to the case of more than two observers and showed that under the ANOVA model, the CCC coincides with the agreement ICC. Hence, this is again a measure of the total agreement. As stated in Section 1, we are interested in measuring the *interobserver* component, i.e., $\sigma_0^2 + \sigma_{S0}^2$ relative to the total (between and within) observer-related variability $\sigma_0^2 + \sigma_{S0}^2 + \sigma_E^2$. Therefore, the new coefficient of interobserver variability (CIV) introduced in this work is $(\sigma_0^2 + \sigma_{S0}^2)/(\sigma_0^2 + \sigma_{S0}^2 + \sigma_E^2)$. The new agreement coefficient is then defined as $1 - CIV$.

Besides the fact that the CIV measures a different parameter than the ICC, there are other important differences between the current and traditional approached to evaluation of observer agreement.

- The CIV has a simple intuitive definition in terms of the difference between the values assigned by different observers to the same subject (see Section 3). The traditional approach uses correlations to evaluate observer agreement.

- The ICC is always defined in terms of variances of the effects in an ANOVA

model. Thus, it is based upon the assumption that the true value of an observer on a given subject is the sum of independent effects attributable to the subject, the observer and the subject-by-observer interaction. It also assumes that the within subject-observer variance (σ_E^2) is the same for all subjects and observers. The approach used in this paper does not make these assumptions.

- The agreement measures based on the traditional approach depend on the between-subjects variability. As we will see in the examples (Section 7), when there is substantial between-subjects variability then the agreement ICC may be very close to unity even when there are important differences between observers. As we pointed out earlier, we assume that the main purpose of the study is to estimate the 'true' value of each subject and hence the between-subjects variability does not affect the CIV.

3. Definition and Estimation of the Coefficient of Interobserver Variability

We denote by Y the variable being observed. In Sections 3-5 we do not make any assumptions regarding the distribution of Y, other than the existence of the second moment. Suppose that there are I subjects and J observers. Each observer makes $K \geq 1$ replicated observations on each subject. Let Y_{ijk} denote the k-th observation ($k = 1, 2, \ldots, K$) made by observer $j (j = 1, 2, \ldots, J)$ on subject $i (i = 1, 2 \ldots, I)$. For a fixed subject-observer combination, these K replicated observations are assumed independent and identically distributed (iid) random variables. By 'independent' we mean that when an observer makes replicated observations on the same subject, then she/he is blinded to her/his previous observations on the same subject.

We will always assume that the subjects are drawn at random from a large population. As to the observers, they may be regarded either as a fixed set or as a random sample from the population of all potential observers. Since we have more than one source of variation, we must indicate the source of variation when we refer to the expectation or variance of a random quantity. We will denote by E_Y the expectation with respect to the distribution of the Y_{ijk}'s (for fixed i, j). The expectation with respect to the subjects' sampling distribution will be denoted E_S, and the expectation with respect to the observers' sampling distribution (in the random observers case) will be denoted E_0. The operator E without a subscript denotes the expectation over all sources of variation. Similar notations will be used for the variances.

Let $\mu_{ij} = E_Y(Y_{ijk})$ and $\sigma_{ij}^2 = Var_Y(Y_{ijk})$. We consider μ_{ij} as the 'true' value that observer j would assign to subject i if s/he could make an infinite number of

replicated observations. The σ_{ij}^2 is the variance of these replicated observation, i.e., it is the *intraobserver* variability. We will use $\cdot$ to denote the arithmetic mean with respect to the relevant index. For quantities that depend only on the μ's and σ's, we will also use $*$ to denote the expectation with respect to the sampling distribution associated with the relevant index. For example $\mu_{I\cdot}$ is always the arithmetic mean of all the μ_{ij}'s for the J observers in the study. In the random observers case we also denote by μ_{i*} the expectation of μ_{ij} over all the possible samples of observers, i.e., $\mu_{i*} = E_0(\mu_{ij})$. We will now consider separately the cases of fixed and random observers.

3.1 Fixed observers

Let $\tau_i^2 = \sum_j (\mu_{ij} - \mu_{i\cdot})^2/(J-1)$ denote the variability among the observers' true values for subject i. The (mean) *interobserver variability* will be defined as $\tau_* = E_S(\tau_i^2)$. In order to define a coefficient that varies between zero and one, we realize that the interobserver variability is a fraction of the total variability associated with the observers. For subject i, the total observer variability is the sum of the variability among the observers' true values, τ_i^2, and the average variability about the true values (the *intraobserver* or error variability), $\sigma_{i\cdot}$. Hence the mean total observer variability is $\tau_*^2 + \sigma_{*\cdot}^2$, where. $\sigma_{*\cdot}^2$ is the mean of $\sigma_{i\cdot}$ over all the subjects. Therefore, we define the *coefficient of interobserver variability* (CIV) as the ratio of the interobserver variability to the total observer variability:

$$\xi = \tau_*^2/(\tau_*^2 + \sigma_{*\cdot}^2) \tag{3.1}$$

Obviously, the CIV is always between 0 and 1, with higher values indicating more variability, i.e., less agreement. A natural coefficient of interobserver agreement is $\psi = 1 - \xi$. One should note that $CIV = 0$, which is equivalent to $\tau_*^2 = 0$, means that there are no true differences among the observers. It does not imply that the observations made by different observers on the same subject must be equal, due to the presence of the intraobserver variances σ_{ij}^2.

It is interesting to note that when there are only two observers, the interobserver variability is

$$\tau_*^2 = \frac{1}{2} E_S(\mu_{i1} - \mu_{i2})^2.$$

When there are more than two observers, it is easy to show that the overall interobserver variability equals to the arithmetic mean of all the pairwise interobserver variabilities.

We now turn to the estimation of CIV. One would expect the observed within-subject variability among the observers' means ($Y_{ij\cdot}$ for $j = 1, 2, \ldots, J$) to

play a major role in this estimation problem. Therefore, let

$$V_i = \frac{1}{J-1}\sum_j (Y_{ij\cdot} - Y_{i\cdot\cdot})^2$$

denote the observed between-observers variability for subject i. Then it is easy to show that $E_Y(V_i) = \tau_i^2 + \sigma_{i\cdot}^2/K$ and $E(V_\cdot) = \tau_*^2 + \sigma_{*\cdot}^2/K$. Estimation of the variances σ_{ij}^2 depends upon whether $K = 1$ or $K \geq 2$. When each observer makes at least two observations on each subject, we define $U_{ij} = \sum_k (Y_{ijk} - Y_{ij\cdot})^2/(K-1)$, so that $E_Y(U_{ij}) = \sigma_{ij}^2$ and $E(U_{\cdot\cdot}) = \sigma_{*\cdot}^2$. Then, an unbiased estimator of the interobserver variability is $\hat{\tau}_*^2 = V_\cdot - U_{\cdot\cdot}/K$ and the CIV is estimated as:

$$\hat{\xi} = \frac{KV_\cdot - U_{\cdot\cdot}}{KV_\cdot + (K-1)U_{\cdot\cdot}} \tag{3.2}$$

When $K = 1$, estimation may be possible under further assumptions. For example, estimation is possible if the μ_{ij}'s follow an additive ANOVA model and the σ_{ij}'s are equal (see Section 6). In general, one can replace U_{ij} by any consistent estimator of σ_{ij}^2.

3.2 Random observers

When we assume that the J observers were selected at random from a larger population, we define $\tau_i^2 = Var_0(\mu_{ij})$ as the interobserver variability for subject i, and $\tau_*^2 = E_S(\tau_i^2)$ as the (mean) interobserver variability. The total observer variability is $E_S(Var_{0.Y}(Y_{ijk})) = \tau_* + \sigma_{**}^2$, and we define the CIV as

$$\xi = \frac{\tau_*^2}{\tau_*^2 + \sigma_{**}^2} \tag{3.3}$$

In order to estimate the CIV in this case, we note that for a fixed subject, V_i depends on the sample of observers. For a fixed subject and a fixed sample of observers, it is easy to show that

$$E_Y(V_i) = \frac{1}{J-1}\sum_j (\mu_{ij} - \mu_{i\cdot})^2 + \frac{1}{K}\sigma_{i\cdot}^2$$

Hence, for a fixed subject

$$E_0(E_Y(V_i)) = Var_0(\mu_{ij}) + \frac{1}{K}\sigma_{i*}^2 = \tau_i^2 + \frac{1}{K}\sigma_{i*}^2.$$

Finally, taking the expectation with respect to the subjects' sampling distribution, we have $E(V_\cdot) = \tau_*^2 + \sigma_{**}/K$. The remaining considerations involving the

estimation of the CIV are almost the same as in the fixed rater case. Thus, when $K \geq 2$ the CIV is again estimated by (3.2).

3.3 Inference on the CIV

In the absence of any distributional assumption, the most obvious inference method is the nonparametric bootstrap. Each of the M bootstrap samples is taken from $\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_I$, where $\tilde{Y}_i$ is the array of the JK observations on subject i, and provides an estimate of the CIV using (3.2). The mean and standard deviation of these M estimates are then used to obtain the bootstrap estimate of the CIV and its standard error. A confidence interval for the CIV can be obtained from the standard error, assuming that the estimator is approximately normally distributed. Alternatively, one can obtain a confidence interval from the percentiles of the empirical distribution of the M estimates. Parametric methods for inference on the CIV can be used when the observations are normally distributed and follow a two-way ANOVA model (see Section 6).

4. The Coefficient of Excess Observer Variability

In this section we introduce an alternative coefficient related to interobserver agreement, the *coefficient of excess observer variability* ($CEOV$), and we show that it is closely related to the CIV. Let Y_{ijk} denote a single observation made by observer j on a randomly selected subject i. Then, we define for each subject

$$W_{ik} = \sum_j (Y_{ijk} - Y_{i\cdot\cdot})^2/(J-1),$$

which is the estimated total observer variability. We then define the $CEOV$ as:

$$\eta = \frac{E_S(W_{ik})}{E_S(W_{ik}|\mu_{i1} = \cdots = \mu_{iJ})} \tag{4.1}$$

The denominator is the expected value of W_{ik} when there is no true interobserver variability, i.e., when $\tau_*^2 = 0$. Thus, the $CEOV$ is the excess variability due to the true differences among the observers. The $CEOV$ varies between 1 and infinity, with $\eta = 1$ indicating no excess observer variability. From Section 3 it follows that in the fixed observers case, $E_S(W_{ik}) = \tau_*^2 + \sigma_{*\cdot}^2$. Therefore

$$\eta = \frac{\tau_*^2 + \sigma_{*\cdot}^2}{\sigma_{*\cdot}^2} = \frac{1}{1-\xi} \ .$$

We see that the $CEOV$ is a one-to-one function of the CIV. The result, which also holds when the observers are random, provides an alternative interpretation

of the CIV. One should also note that $1/\eta = \psi$, which is the coefficient of interobserver agreement proposed in Section 3.

In the case of random observers, the $CEOV$ can also be interpreted in the context of estimation. Suppose that $K = 1$ and we are interested in estimating the true value, μ_{i*}, of each subject by $Y_{i\cdot}$. Then the mean (over subjects) variance of this estimator is $(\tau_*^2 + \sigma_{**}^2)/J$. Therefore, the $CEOV$ is the excess in the estimation variance due to the interobserver variability. (One should note that the interobserver variability does not depend on the number of observers used in the study).

5. An Aplication in Study Design

In the previous sections we considered data from an agreement study, which is a study designed for the purpose of evaluating the agreement among observers. In many cases, an agreement study is followed by a more comprehensive study, a measurement study, in which investigators are interested in estimating the true value of Y for each subject. The measurement studies usually involve a limited number of observers, as well as a smaller number of replications (or no replications). In this section we show how the results from an agreement study can be used to help in the design of a measurement study when it is assumed that both studies are based on samples from the same population. We confine the discussion to the case of random observers.

We define the true value of Y on subject i, μ_{i*}, as the mean of all the true evaluations μ_{ij} that could be made if we had an infinite number of observers. If the measurement study involves J observers and each observer makes K measurements of each subject, then $Y_{i\cdot\cdot}$ is an unbiased estimator of μ_{i*}. Then the squared error for subject i, i.e., the expectation of squared distance between the estimated and the true value, is: $Var(Y_{i\cdot\cdot}) = (\tau_i^2 + \sigma_{i*}^2/K)/J$. The mean squared error (over all the subjects) is $D_{JK}^2 = (\tau_*^2 + \sigma_{**}/K)/J$. Suppose that the investigator plans to make a total of M measurements per subject, and has to choose between two designs: (i) each of M observers makes one measurement on each subject, or (ii) a single observer makes M measurements on each subject. Then the ratio of the mean squared errors for these designs is:

$$\frac{D_{M1}^2}{D_{1M}^2} = \frac{1}{1 + (M-1)\xi}$$

Thus, design (i) is always more efficient (i.e., involves a smaller error) than design (ii), and the relative efficiency of design (i) compared to design (ii) (the reciprocal of the above ratio) is an increasing linear function of the CIV.

6. Interobserver Agreement in the Two-way ANOVA Model

In this section we assume that the observations follow a two-way ANOVA model and we write the estimator (3.2) of the CIV is terms of the ANOVA sums of squares. We also show that the F statistic for testing the hypothesis $CIV = 0$ (under the assumption of normality) is a simple function of the CIV estimator.

6.1 Fixed observers

We first assume that $K \geq 2$. In this case the ANOVA model will include a term for the subject by observer interaction. Thus, the appropriate model is a 2-way mixed model with interaction:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} = \mu_{ij} + \epsilon_{ijk} \tag{6.1}$$

where the α's, γ's and ϵ's are iid with $E(\alpha_i) = E(\gamma_{ij}) = E(\epsilon_{ijk}) = 0$, $Var(\alpha_i) = \sigma_S^2, Var(\gamma_{ij}) = \sigma_{S0}^2, Var(\epsilon_{ijk}) = \sigma_E^2$. The parameters $\beta_1, \ldots, \beta_J$ are the fixed observer effects with $\beta_\cdot = 0$. Let $s_0^2 = \sum_j \beta_j^2/(J-1)$. Then, it is easy to see that for subject i:

$$\tau_i^2 = s_0^2 + \frac{1}{J-1}\sum_j (\gamma_{ij} - \gamma_{i\cdot})^2 + \frac{2}{J-1}\sum_j \beta_j(\gamma_{ij} - \gamma_{i\cdot})$$

Taking the expectation with respect to subjects, the interobserver variability is $\tau_*^2 = s_0^2 + \sigma_{S0}^2$ Noting that $\sigma_{*\cdot}^2 = \sigma_E^2$, the CIV is

$$\xi = \frac{s_0^2 + \sigma_{S0}^2}{s_0^2 + \sigma_{S0}^2 + \sigma_E^2}.$$

The estimator (3.2) of the CIV can be derived from the usual ANOVA sums of squares. Let

$$SSBOWS = K\sum_i \sum_j (Y_{ij\cdot} - Y_{i\cdot\cdot})^2$$

be the sum of squares between observers within subjects. It is easy to show that this is the sum of the sum of squares between observers and the sum of squares for interaction. The corresponding mean square, $MSBOWS = SSBOWS/I(J-1)$, equals to $KV_\cdot$ in the notation of Section 3. Also, $U_{\cdot\cdot} = MSE$, the mean squares for error. Therefore:

$$\hat{\xi} = \frac{MSBOWS - MSE}{MSBOWS + (K-1)MSE} \tag{6.2}$$

The equations (3.2) and (6.2) always produce the same value of the estimated CIV.

When $K = 1$, we have to assume that there is no subject by observer interaction ($\sigma^2_{S0} = 0$), and modify the expression for CIV accordingly. In this case, the sum of squares for error (SSE) is calculated as the sum of the squares of the residuals from the additive model, and the error variance (σ^2_E) is estimated by the corresponding mean square for error (MSE). Also, $SSBOWS$ can be obtained as the sum of the sum of squares between observers and the sum of squares for error, and the corresponding mean square equals $V.$. Hence, $\hat{\xi} = 1 - MSE/MSBOWS$. (This is identical to the expression (6.2) with $K = 1$, but one must remember that MSE is defined differently when $K = 1$.)

6.2 Random observers

When $K \geq 2$, the two way model (6.1) becomes a random effects model where β_j is a random variable with mean 0 and variance σ^2_0 Then

$$\tau^2_* = \tau^2_i = Var_0(\mu_{ij}) = \sigma^2_0 + \sigma^2_{S0}, \quad \sigma^2_{**} = \sigma^2_E,$$

and the CIV is defined as:

$$\xi = \frac{\sigma^2_0 + \sigma^2_{S0}}{\sigma^2_0 + \sigma^2_{S0} + \sigma^2_E}$$

All the arguments we made in deriving the estimator (6.2) of the CIV in the fixed observers case remain valid when the observers are random. Similarly, when $K = 1$ and there is no subject by observer interaction, the estimator of CIV in the random observers case is the same as in the fixed raters case.

6.3 Testing the hypothesis $CIV = 0$

In the setting dealt with in the work, one may wish to test the hypothesis of no interobserver variability, i.e., $CIV = 0$. This can be easily done when we assume that the observations follow the two-way ANOVA model, all the random effects are normally distributed and the usual independence requirements are satisfied.

When $K = 1$, one can use the standard ANOVA methods to test the hypothesis $s_0 = 0$ (for fixed observers) or $\sigma_0 = 0$ (for random observers). When $K \geq 2$ and the observers are considered fixed, then the hypothesis of interest is $H_0 : \ s^2_0 = \sigma^2_{S0} = 0$. It is easy to see that

$$E(MSBOWS) = K(s^2_0 + \sigma^2_{S0}) + \sigma^2_E.$$

Also, the corresponding sum of squares is the sum of the sum of squares between observers and the sum of squares for interaction, each of which has an independent $\sigma^2_E \chi^2$ distribution under H_0. Hence under H_0, $MSBOWS$ is distributed as a

$\sigma_E^2\chi^2$ divided by its number of degrees of freedom and is independent of the MSE. This implies that an F statistic for H_0 is:

$$F = \frac{MSBOWS}{MSE} = \frac{1+(K-1)\hat{\xi}}{1-\hat{\xi}}.$$

The corresponding degrees of freedom are $I(J-1)$ and $IJ(K-1)$. All the above considerations remain valid in the random observers case when s_0^2 is replaced by σ_0^2.

7. Examples

Example 7.1 Coronary artery calcium scoring gives an indication as to the presence or absence of coronary artery arteriosclerosis and its severity. The presence of CT-detected calcium indicates the presence of underlying arteriosclerosis. The actual score, which is known as AJ130, is based on the area of the calcified plaque multiplied by a weighting factor which depends on the highest density in the area of the plaque. Various software programs use the AJ130, and the score is applied by drawing rectangles around operator-selected regions. Each vessel is scored independently, and a total is given.

The data used for this example are the total AJ130 scores from 12 patients. Each patient was scored twice by each of two radiologists. The radiologists are labeled A and B, and the replications are labeled 1 and 2. Table 1 presents the four scores for each patient. We see that there is considerable variability among subjects but the within-subject variability is relatively small.

Table 1: Calcium scores on 12 patients

	Patients											
	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

The agreement ICC and the CCC for these data are both equal to 0.997, which means that there is practically a perfect agreement between the two radiologists. On the other hand, the CIV is 0.246, hence the new coefficient of interobserver agreement is $\hat{\psi} = 0.754$. The CEOV equals 1.33, which means that the total observer variability is 33% higher than one would expect if there were no differences between the two radiologists. The data indeed suggests that the

agreement between the two radiologists is less than perfect, and that the values of the ICC and the CCC are inflated as a result of the large between-subjects variability. This example demonstrates that the ICC and the CCC are unable to reflect observer disagreement when the between-subject variability is substantially higher than the within-subject variability.

Example 7.2 The concepts and methods developed in this work can be used in comparing different methods or instruments for measuring a certain quantity on a sample of study subjects, rather than comparing different human observers. Eliasziw *et al.* (1994) presented data from a study conducted to compare a large universal manual goniometer and a Lamoreux-type electrogoniometer for measuring the knee joint angle. Twenty-nine subjects were measured three consecutive times on each of the two goniometers. The measurements ranged from -14 to $+19$ degrees, with the means for the two methods calculated as 1.44 and 0.05, respectively. The point estimate (3.2) of the CIV is 0.713, and the percentile-based bootstrap 95% confidence interval (with 1000 replications) is [0.571, 0.824]. The estimated interobserver agreement coefficient is $\hat{\psi} = 1 - 0.713 = 0.287$ and the estimate of the $CEOV$ is 3.48, i.e., the total observer variability is almost 3.5 times higher than one would expect if the two goniometers were equivalent. Using the ANOVA notations, the following sums of squares were obtained: between subjects, 8757.862, between observers (goniometers), 84.144, interaction, 126.023 and error, 99.333. From these sums of squares we calculated $SSBOWS = 84.144 + 126.023 = 210.167$, $MSBOWS = 210.167/29 = 7.247$, $MSE = 99.333/116 = 0.856$. These mean squares yield an estimated CIV of $(7.247 - 0.856)/(7.247 + 2 \times 0.856) = 0.713$. If we assume normality, then the F statistic for testing $CIV = 0$ is $7.247/0.856 = 8.463$, with 29 and 116 degrees of freedom ($p < 0.001$). This indicates that there is a substantial variability between the two goniometers. For comparison, the value of the ICC reported by Eliasziw *et al.* is 0.961.

8. Discussion

This work presents a new index for assessing the agreement between observers. It is based on the squared distances between the true values assigned by different observers to the same subject, i.e., the interobserver variability. The CIV measures the interobserver component of the total disagreement between observers, while existing indices of observer agreement measure the total (inter + intra) observer-related disagreement. In other words, our approach distinguishes between true disagreements among observers and the variability resulting from the differences among (real or hypothetical) observations made by the same observer on the same subject. This simple idea of separating the total observer

disagreement to components representing the between and within observer disagreements has not received much attention in the literature. Recently, Barnhart *et al.* (2004) used the concept of interobserver variability to define an interobserver version of the CCC as a measure of agreement between observers who make replicated measurements on each subject. The approach used in the present work is more general than the traditional approach as we do not assume that the observations follow an ANOVA model. The new approach also seems more intuitive and there is a simple relationship between the interobserver variability among several observers and the mean of all the pairwise differences.

The ICC and CCC compare the (total) observer variability to the within-subject variability. As we have seen in the examples, these coefficients may produce unreasonably high estimates of observer agreement when there is substantial variability among study subjects. On the other hand, the CIV compares the between-observers variability to the total observer-related variability. In other words, it uses the error variance, which is based upon the within subject-observer variabilities as reference. This is consistent with the common approach used in statistical inference. The error variance is most precisely estimated when each observer makes replicated observations on each subject. When using replicated observations, one must ensure to the extent possible that they are independent for a given subject-observer combination in the sense that the observer should not be able to recall her/his previous readings for the same subject. Estimation of the CIV when there are no replications requires additional assumptions, such as the assumption of additivity in the 2-way ANOVA model.

We also introduced the concept of excess observer variability. The $CEOV$, which is a simple function of the CIV, compares the actual total observer variability to what one would expect if the observers were equivalent. The new coefficient of interobserver agreement (ψ) is the reciprocal of the $CEOV$. In addition, we showed how the CIV and the $CEOV$ are related to the estimation variance in studies designed to estimate the true value of each study subject.

In this work we focused on evaluating the agreement between different observers, or measurements methods. The same concepts can be used to evaluate the agreement between observers and a known gold standard. For example, if we want to compare the measurements of a single observer with the corresponding true values, we can set $J = 2$, define Y_{i1k} for each k as the true value of subject i, and define Y_{i2k} as the k-th measurement of the observer. In this case it is natural to assume $\sigma_{i1} = 0$ for all i. The interobserver variability for subject i, τ_i^2, is one-half the squared difference between the true value of this subject ($Y_{i1k} = \mu_{i1}$) and the 'true value' of the observer on this subject (μ_{i2}). The overall observer variability, τ_*^2, is the mean over all the subjects.

Future Research

We plan to explore generalizations of the concepts and methods developed in this work. Possible extensions include: (a) using methods based on generalized estimation equations (GEE) and on mixed linear models for inference; (b) unequal number of replications and missing observations; (c) more complicated designs, such as having different subsets of observers evaluating different subjects, or multifactor studies were each observer uses each of several measuring methods to evaluate each subject.

Acknowledgement

We would like to thank the Editor and a referee for helpful comments. This research was supported in part by the University Research Committee of Emory University.

References

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

Barnhart, H. X., Song, J. and Haber, M. (2004). Assessing agreement in studies designed with replicated readings. *Statistics in Medicine*, (in print).

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

Eliasziw, M., Young, S. L., Woodbury, M. G. and Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using goniometric measurements as an example. *Physical Therapy* **74**, 777-788.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd Ltd.

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

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods* **1**, 30-46.

Received June 14, 2003; accepted September 6, 2003.

Michael Haber
Department of Biostatistics
Rollins School of Public Health
Emory University, 1518 Clifton Rd. N.E.
Atlanta, GA 30322, USA
mhaber@sph.emory.edu

Huiman X. Barnhart
Department of Biostatistics and Bioinformatics
Duke University Medical Center
Durham, NC, USA

Jingli Song
Eli Lilly and Company
Indianapolis, IN, USA

James Gruden
Department of Radiology
Emory University School of Medicine
Atlanta, GA, USA