STATISTICS IN MEDICINE, VOL. 13, 737-745 (1994)

GENERAL METHODOLOGY II

MEASURES OF AGREEMENT: A SINGLE PROCEDURE

JOHN J. BARTKO

Division of Epidemiology and Services Research, National Institute of Mental Health, NIH Campus, Building 10, Room 3N-204, Bethesda, Maryland, 20892, U.S.A.

SUMMARY

An assessment of measurement agreement made by devices, laboratories, or raters is important in medical practice and research. The setting in which each randomly selected subject is rated by the same two raters raises assorted questions regarding rater agreement. The intraclass correlation coefficient (ICC) is one measure of reliability. The paired *t*-test can be used to evaluate the overall ratings or bias of the two raters, while their variances can be assessed with Pitman's test. The Bradley-Blackwood test can be used for a simultaneous test of their means and variances. A single method that provides results for these approaches is proposed and the bivariate confidence ellipse is suggested to provide boundaries for dispersion.

INTRODUCTION

There is often a requirement in medicine, as in other sciences, to assess the agreement of two measurements made by devices, laboratories or raters. Devices can measure clinical signs and symptoms, laboratories analyse fluids, materials, and tissues, while raters and judges usually evaluate patients or subjects. Various questions regarding agreement can be posed. How much bias and variance is there between the two measurements? Can the measuring instruments be used interchangeably? What is the overall level of agreement? How does one define and measure agreement?

For continuous data, one measure of reliability agreement is the intraclass correlation coefficient (ICC) (Bartko[1]), which is estimated using variance components from appropriate analysis of variance models. Since measures of reliability agreement are variance dependent, the ICC can be low even when a high level of agreement is evident among raters. The most extreme example is where there is 100 per cent rater agreement, but zero between-subject variance. In general, low ICC's occur when the variation between subjects is low relative to that within-subjects or the within-subjects variance is large relative to the between-subjects variance. Either situation can yield a low *F* statistic, the test for the ICC. The size of the ICC is related to the size of the *F* statistic of the mean square between subjects to mean square error. In this paper the term subjects is used in the generic sense to represent the experimental unit, while raters is used for the measuring device.

For the same degree of within-subject variance, the greater the between-subject variance, the greater the ICC. It is this property of the ICC which prompts criticism from Bland and Altman,[2] who prefer a method of assessing agreement that is not dependent on the nature of the group of subjects chosen for the study. Taking another view, Shrout *et al.*[3] point out the distinction between discarding an agreement statistic and the very real problem of making distinctions in homogeneous settings. The term 'homogeneous settings' reflects the case where there is little

Table I. Data format for agreement exercises for two raters and N subjects

Subject	Rater 1 $X1$	Rater 2 $X2$	Differences d
1	$X(1,1)$	$X(1,2)$	$X(1,1) - X(1,2)$
2	$X(2,1)$	$X(2,2)$	$X(2,1) - X(2,2)$
3	$X(3,1)$	$X(3,2)$	$X(3,1) - X(3,2)$
•	•	•	• •
•	•	•	• •
i	$X(i,1)$	$X(i,2)$	$X(i,1) - X(i,2)$
•	•	•	• •
•	•	•	• •
N	$X(N,1)$	$X(N,2)$	$X(N,1) - X(N,2)$

	Parmeters		Statistics	
Means	μ_1	μ_2	$\bar{X}_1$	$\bar{X}_2$
Variances	σ_1^2	σ_2^2	s_1^2	s_2^2
Correlation	$\rho(X1,X2)$		$r(X1,X2)$	

where $i = 1, 2, \ldots, N$, a random sample of subjects and $j = 1, 2$, the number of fixed effect raters.

subject to subject variation (low between-subjects variance) in ratings. If the between-subject variance is low (even with low within-subject variance), measures of reliability are problematic, not from an arithmetic point of view but from a design one. Such settings offer little opportunity for the raters to use the full range of a scale or measurement system. Thus, whatever their degree of agreement, it lies only within a narrow segment of the scale and the reliable use of the full scale is left unanswered. In such design settings, where low reliability values result, some investigators are quick to call for supplemental statistical measures. Low reliability values may instead serve to portend that a variable is not sufficiently precise in definition or administration to warrant continued use. In such settings the ICC often provides opportunities for statistical and scientific discovery.

It is axiomatic that statistics should be purposeful and that there is no universal statistic. The ICC is a measure of reliability agreement, but there are other definitions of agreement. The paired t-test can be used to evaluate the overall ratings or bias of two raters. Their variances can be assessed with Pitman's test.[4] A simultaneous test of the two means and variances can be performed using the Bradley and Blackwood test.[5] Russell *et al.*[6] have explored various ICC simulations with this procedure. Graphic procedures that plot sums and difference provide important variance information. This paper proposes a single method that provides results for these procedures.

METHODS

We assume normally distributed data, where each of N randomly selected subjects is rated by each of the same two raters who are the only raters of interest. These assumptions lead us naturally to the two-way ANOVA mixed model. The data format is presented in Table I and the two-way mixed model is presented in Table II.

Table II. Two-way mixed model ANOVA for ICC for $k = 2$ raters

Source	d.f.	MS	Expected mean squares
Between-subjects	$N - 1$	MSB	$\sigma_e^2 + k\sigma_p^2$
Within-subjects	N	MSW	$\sigma_e^2 + f\sigma_1^2 + \Sigma r^2/(k - 1)$
Between raters	1	MSR	$\sigma_e^2 + f\sigma_1^2 + N\Sigma r^2/(k - 1)$
Residual	$N - 1$	MSE	$\sigma_e^2 + f\sigma_1^2$
Total	$2N - 1$		

(where in general $f = k/(k - 1)$.

The two-way mixed model can be written: $x_{ij} = \mu + p_i + r_j + (pr)_{ij} + \varepsilon_{ij}$ where μ is the overall effect common to all observations; p_i is a random variable common to the ith subject, r_j is a fixed effect common to the jth rater, $(pr)_{ij}$ is the interaction term, a random variable for observation (i,j), and ε_{ij} is the error associated with observation (i,j). The usual constraints and assumptions are: $\Sigma_j(pr)_{ij} = 0$ and $(pr)_{ij} = N[0, \sigma_1^2(k - 1)/k]$. A consequence of these assumptions is that the within-subject $\text{cov}(x_{ij}, x_{is}) = \sigma_p^2 - (\sigma_1^2)/(k - 1)$.

The population intraclass correlation coefficient is defined to be: $\rho(\text{ICC}) = [\sigma_p^2 - (\sigma_1^2)/(k - 1)]/(\sigma_e^2 + \sigma_1^2 + \sigma_p^2)$ and is estimated by: $\text{ICC} = (MSB - MSE)/[MSB + (k - 1)MSE]$. In particular, since the number of raters (k) is two, this is simply:

$$\text{ICC} = (MSB - MSE)/(MSB + MSE).$$

The ICC ranges from $[-(1/(k - 1))$ to $+1]$ or $(-1$ to $+1)$ where $k = 2$.

The methods presented for discussing agreement and reliability include the Bradley–Blackwood simultaneous test on two means and variances, the intraclass correlation coefficient (ICC), Student's paired t-test, Pitman's test on correlated variances, and ellipse graphics.

SINGLE PROCEDURE FOR SEVERAL MEASURES OF AGREEMENT

A single procedure that provides values for each of the above mentioned agreement methods springs from the Bradley and Blackwood[5] method of simultaneously testing means and variances for paired data.

Bradley–Blackwood procedure

Let $y = (X_1 - X_2)$, and $x = (X_1 + X_2)/2$. The Bradley–Blackwood procedure fits y on x, that is

$$y = \beta_0 + \beta_1 x \qquad (1)$$

where β_0 and β_1 the intercept and slope, respectively, are given by

$$\beta_0 = (\mu_1 - \mu_2) - 0{\cdot}5[(\sigma_1^2 - \sigma_2^2)/\sigma_x^2][(\mu_1 + \mu_2)/2]$$

and

$$\beta_1 = 0{\cdot}5(\sigma_1^2 - \sigma_2^2)/\sigma_x^2.$$

Now $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ if, and only if, $\beta_0 = \beta_1 = 0$, so a simultaneous test of the equivalence of the means and variances of X_1 and X_2 is an F test, calculated from the results of a regression of y on x. The test (Bradley–Blackwood[5]) statistic is

$$F(2, N - 2) = (\Sigma y^2 - SSReg)/(2\,MSReg) \qquad (2)$$

Table III. Regression and two rater two-way mixed model identities

Source	d.f.	SS	Regression identity
Between-subjects	$N-1$	SSB	$2Sxx$
Within-subjects	N	SSW	$0{\cdot}5\Sigma y^2$
Between raters	1	SSR	$(\Sigma y)^2/(2N)$
Residual	$N-1$	SSE	$0{\cdot}5Syy$
Total	$2N-1$		

where $y = (X_1 - X_2)$, $x = (X_1 + X_2)/2$, $Sxx = \Sigma x^2 - (\Sigma x)^2/N$ and $Syy = \Sigma y^2 - (\Sigma y)^2/N$. For the ellipse: σ_x^2 is estimated by $Sxx/(N-1) = 0{\cdot}5MSB$ and σ_y^2 is estimated by $Syy/(N-1) = 2MSE$.

where *SSReg* and *MSReg* are the residual sum of squares and the mean square with $N-2$ degrees of freedom, respectively, from the regression of y on x. Table III provides regression and two-way mixed model ANOVA identities.

From the identities expressed in Table III and from the usual basic statistics, that are an integral part of any simple linear regression output of y on x, the other 'agreement' procedures can be written as:

Intraclass correlation coefficient (ICC)

For two fixed raters, assuming a two-way mixed model ANOVA, the details of which appear in Table II, the mixed model ICC is defined to be:[7-9]

$$\text{ICC} = (MSB - MSE)/(MSB + MSE).$$

or equivalently

$$\text{ICC} = (4Sxx - Syy)/(4Sxx + Syy). \tag{3}$$

Since the ICC may not be as familiar as the other mentioned procedures, its F test and confidence interval are presented. The ICC is tested against zero by $F_0 = MSB/MSE$ with $(N-1)$ and $(N-1)$ degrees of freedom. For the confidence interval, let $F_L = F_0/[F_{1-\alpha/2}; (N-1), (N-1)]$ and $F_U = F_0[F_{1-\alpha/2}; (N-1), (N-1)]]$ then

$$(F_L - 1)/[F_L + 1] < \rho(\text{ICC}) < (F_U - 1)/[F_U + 1] \tag{4}$$

is the $(1-\alpha)$ 100 per cent confidence interval for ρ(ICC).[9] As an aside the ICC can also be written as $(F_0 - 1)/(F_0 + 1)$.

Student's *t*-test on rater means

The paired t-test can be used to assess the mean levels of response (bias) between the two raters. The test of H_0: $\mu_1 - \mu_2 = 0$, against the alternative H_1: $\mu_1 - \mu_2 \neq 0$, is:

$$t(N-1) = \bar{d}/[Sdd/(N(N-1))]^{1/2}$$

where $d_i = X_{1i} - X_{2i}$, $i = 1$ to N. $Sdd = \Sigma d^2 - (\Sigma d)^2/N$ and $\bar{d} = \Sigma d/N$, or equivalently:

$$t(N-1) = (\Sigma y/N)/[Syy/(N(N-1))]^{1/2}. \tag{5}$$

Pitman's test on correlated variances

The test procedure for H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$ is:

$$t(N-2) = (F-1)(N-2)^{1/2}/[(1-r^2(X1, X2))4F]^{1/2} \tag{6}$$

where $F = \sigma_1^2/\sigma_2^2$ is replaced by respective sample estimates.[10]

Pitman's test, for equal rater (correlated) variances, is identical to the test of the slope equal to zero in the regression of y on x. The proof, not presented here, is a straightforward manipulation of algebra.

Pitman's t-test of slope or the test that the variances are equal is,

$$t(N-2) = (\beta_1)/(MSReg/Sxx)^{1/2} = (Sxy/Sxx)/(MSReg/Sxx)^{1/2}. \tag{7}$$

Note that the Pitman test and Bradley and Blackwood use $(X_1 + X_2)$, not the average. For the results presented here, the sum versus the average produces identical results, although the slope with the average x is twice the slope on the sums of x. The t statistics on zero slope is the same and is identical to Pitman's test on correlated variances. The Pearson correlation between x and y is by definition the same as Pitman's correlation on sums and differences. The Bradley–Blackwood result is the same.

Graphics ellipse

Altman and Bland,[11] in a similar setting dealing with pairs of observations, suggests plotting the within-subject differences $(X_1 - X_2)$ on the ordinate versus the average response per subject $(X_1 + X_2)/2$, the between-subject differences, on the abscissa. This suggestion is expanded upon in this paper by advancing the bivariate confidence ellipse to amplify dispersion. The equation of the ellipse is:

$$(x-\bar{x})^2/\sigma_x^2 - 2\rho(x-\bar{x})(y-\bar{y})/\sigma_x\sigma_y + (y-\bar{y})^2/\sigma_y^2 = \chi^2(2\text{df})(1-\rho^2) \tag{8}$$

where $y = (X_1 - X_2)$ and $x = (X_1 + X_2)/2$. Also, $\sigma_x^2 = \sigma^2[(X_1 + X_2)/2]$ is estimated using $0{\cdot}5\,MSB = Sxx/(N-1)$ (see Table III); $\sigma_y^2 = \sigma^2(X_1 - X_2)$ is estimated using $2\,MSE = Syy/(N-1)$, and $\rho = \text{corr}(x, y)$ is estimated from the regression of y on x. Pearson's correlation of (x, y) is the same as Pitman's correlation of sums and differences. Techniques for plotting an ellipse can be found in Altman,[12] who uses chi square not the F statistic in the equation of the ellipse.

RESULTS

The above procedures are illustrated with a data set (Table IV) taken from one of 96 real data eye tracking studies where responses are measured in milliseconds.

From the data and basic statistics in Table IV we can find the following:

Bradely–Blackwood simultaneous test on means and variances
From (2), $F(2,7) = (119 - 112{\cdot}67)/(2 \times 16{\cdot}10) = 0{\cdot}20$.

Mixed Model Intraclass Correlation Coefficient
From (3) the ICC $= (4 \times 40{\cdot}5 - 116{\cdot}23)/(4 \times 40{\cdot}5 + 116{\cdot}23) = 0{\cdot}16$.
The F test on the ICC is $F(N-1, N-1) = MSB/MSE$. Thus, $F(8,8) = 2\,Sxx/0{\cdot}5\,Syy = 81/58{\cdot}115 = 104$ (From Table III and Table IV). Since the d.f. on both numerator and denominator sums of squares for the mixed model are $N-1$, the test is written with sums of squares. From (4) the 95 per cent confidence interval is $-0{\cdot}52 < \rho(\text{ICC}) < 0{\cdot}72$.

Tabe IV. Eye tracking data in milliseconds for nine elderly men

	Raters			
Subject	$X1$	$X2$	$y = (X1 - X2)$	$x = (X1 + X2)/2$
1	52	58	− 6	55
2	53	55	− 2	54
3	59	56	3	57·5
4	60	54	6	57
5	59	59	0	59
6	59	60	− 1	59·5
7	57	59	− 2	58
8	53	58	− 5	55·5
9	54	52	2	53
N	9	9		
Σ	506	511	− 5	508·5
$\bar{x}$	56·2	56·8	− 0·56	56·5
Σ^2			119	28770·75
Syy and Sxx			116·22	40·5
s^2	10·19	7·19	14·53	5·0625
$sd = s$	3·2	2·7	3·81	2·25
Regression of y on x: $y = -17{\cdot}3 + 0{\cdot}3x$ $r(x, y) = 0{\cdot}175$				

ANOVA regression Source	SS	d.f.	MS	F
Due to regression	3·56	1	3.56	0·22
About regression	112·67	7	16·10	(= $MSReg$)
Total Syy =	116·23	8		

Two way ANOVA Source	d.f.	SS	MS	F	ICC
Between-subjects	8	81	10·125	1·39	0·16
Within-subjects	9	59·5	6·611		
Between Raters	1	1·39	1·39	0·19	
Residual	8	58·11	7·26		
Total	17				

Student's paired t on means

From (5), $t(8) = -5/9/[116{\cdot}23/(9 \times 8)]^{1/2} = -0{\cdot}44$.

Pitman's test on correlated variances

From (7), $t(7) = 0{\cdot}3/(16{\cdot}10/40{\cdot}5)^{1/2} = 0{\cdot}48$.

Ellipse

For the ellipse (Figure 1), estimates for σ_x^2 and σ_y^2 are, respectively, $40{\cdot}5/8 = 5{\cdot}06$ and $116{\cdot}23/8 = 14{\cdot}53$; $r(x, y) = (3{\cdot}56/116/23)^{1/2} = 0{\cdot}175$. For a 95 per cent confidence ellipse, the chi square table value is 5·99. Thus from (8):

$$(x - 56{\cdot}5)^2/5{\cdot}06 - 2(0{\cdot}175)(x - 56{\cdot}5)(y + 0{\cdot}56)/(2{\cdot}25 \times 3{\cdot}81) + (y + 0{\cdot}56)^2/14{\cdot}53 = 5{\cdot}99(1 - 0{\cdot}175^2).$$

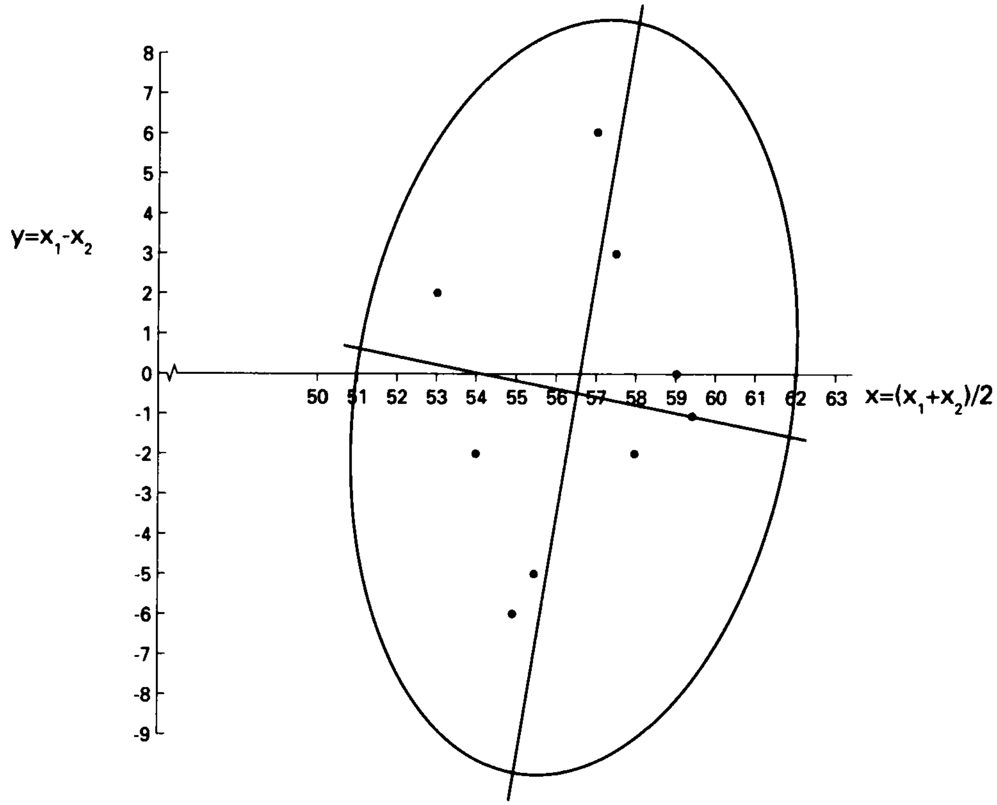

Figure 1. 95 per cent confidence ellipse for the data in Table IV, where the *x* axis is regarded as the between-subjects variance and the *y* axis as the within-subjects variance

Note that the ellipse clearly illustrates the smaller between-subject variation relative to the error mean square. The minor axis, which is related to the variance between subjects, is small relative to the major axis, which in turn is related to the error mean square.

The ICC is low, as are the paired *t*, Pitman's test and the Bradley–Blackwood test. The ellipse's pronounced vertical orientation reflects the low between-subjects variation relative to the larger within-subjects variation and thus the resulting low *F* statistic.

DISCUSSION

In the early (pre-statistician) stages of this study, where 96 real data sets similar to the above (Table IV) were investigated, the paired *t*-test was used to assess agreement. The *t*-test, while useful in assessing the overall ratings (bias) of two raters, provides no guidance apropos individual within-subject agreement. A non-significant *t* does not assure good individual within-subject agreement. In fact a *t* of zero can produce a negative ICC, as it would, for example, with the data set (1, 5), (2, 4), (3, 3), (4, 2), (5, 1).

On the other hand, a statistically significant paired *t* does not guarantee a low ICC. It is possible to have a high ICC, say 0·75 or greater, and a statistically significant paired *t*-test as would be the case with the data set (1, 1·1), (2, 2·5), (3, 3·3), (4, 4·2), (5, 5·5) where the paired *t* is 4,

$p < 0{\cdot}02$ with an effect size of 1·8. The ICC is 0·98. The ICC is a function of both the MSB and MSE, while the paired t-test is a function of the MSE only, which if sufficiently small relative to its numerator can produce a statistically significant value. In nine of the 96 real research eye tracking data sets, the ICC was greater than 0·75 while the paired t was statistically significant. In four of these cases the effect sizes were between 0·7 and 0·8, while in the other five cases the effect sizes were greater than 1·0.

The ICC was also useful in revealing a low F ratio of between-subject variance to the within-subject variance in the elderly men. In the three other groups (there were four subject groups in the study that led to this article), elderly women, young men and young women, the F ratio of MSB to MSE ranged from 8 to 16 and the ICC's from 0·78 to 0·88.

For a quick appreciation of data, the mixed model ANOVA (Table IV) produces the ICC as well as the F test equivalent of the paired t on raters.

A two-way random model also can be assumed for the ICC. It was not in this study since the two raters were the only raters of interest and were regarded as fixed effects. With the random model, the k raters are assumed to be a random samples from a larger population of raters, where each rater judges each subject. For a two rater random model the algorithms presented above apply except for the ICC which is estimated differently.[7,9] Testing of means is natural with the mixed model. However, with a random model, the focus of statistical inference is on variance components, not on specific means.

In the plot of x against y, the x axis reflects the difference between subjects and the y axis reflects the differences between ratings. In typical reliability studies it is desirable to have low differences between readings for a given subject and large differences between subjects. Large between-subject differences usually reflect the condition where the raters are given a thorough opportunity to test their skills with the full range of a measurement scale. Using only normal subjects or subjects with predefined severity is an example of a narrow range study, where rater reliability is a test only of agreement on the absence of disease or on a subsection of the (disease) scale, respectively, not a test of the full instrument.

The ellipse provides visual guidelines for the scatter plot and for focus on outlier responses. It is a convenient way to appreciate the within- and between-subject variances. The major and minor axis of the ellipse are related to the size of the variances of x and y, with its orientation relative to the sizes of the variances. For σ_x^2 greater than σ_y^2 its orientation is horizontal. For σ_x^2 less than σ_y^2 its orientation is vertical (Figure 1). The more horizontally oriented the ellipse, the greater the ICC. The ICC is greater than 0·5 as σ_x^2 is greater than $0{\cdot}75\sigma_y^2$ or as the F statistic is greater than 3.

As suggested by the reviewers, another example may serve to illustrate the differential application of these procedures. In studies of monozygotic twins discordant for schizophrenia it is not uncommon for response levels of particular variables to be similar in the twin groups, but with variances higher in the ill group compared to the healthy control one. The levels question can be addressed by the paired t. Pitman's test addresses the disparity in variances. For an overall type I error preservation the Bradley–Blackwood procedure can be used. The (low) ICC will support the disparity of measurements within the twin pairs. One general advantage and features of the ICC is that as an indexed statistic, its values can be mapped and thereby perhaps more appreciated, to the within variance. The x, y scatter plot of the data and the ellipse are useful, since the differences in the twin pairs (y), will be large relative to their averages (x). The ellipse will be vertically oriented and its presentation is a tool of insight for the investigators.

A single procedure that yields results for several approaches to reliability and agreement has been presented. It is worth noting that Pearson's correlation which measures the strength of linear association, and thus not agreement between two variables, is not among them.[1,13] It should not be interpreted that these procedures for exploring agreement are required or even suitable in all instances. Research decisions should be based upon the nature of the study and the purposes of the various agreement measures.

ACKNOWLEDGEMENTS

Appreciation is expressed to the editors and reviewers for their constructive comments and thoughtful suggestions for improvement.

REFERENCES

1. Bartko, J. J. and Carpenter, W. T. Jr. 'On the methods and theory of reliability, *The Journal of Nervous and Mental Disease*, **163**, 307–317 (1976).
2. Bland, J. M. and Altman, D. G. 'A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement', *Computations in Biology and Medicine*, **20**, 337–340 (1990).
3. Shrout, P. E., Spitzer, R. L. and Fleiss, J. L. 'Quantification of agreement in psychiatric diagnosis revisited', *Archives of General Psychiatry*, **44**, 172–177 (1987).
4. Pitman, E. J. G. 'A note on the normal correlation', *Biometrika*, **31**, 9–12 (1939).
5. Bradley, E. L. and Blackwood, L. G. 'Comparing paired data: A simultaneous test of means and variances', *The American Statistician*, **43**, 234–235 (1989).
6. Russell, C. M., Williamson, D. F., Bartko, J. J. and Bradley, E. L. 'A simulation study of reliability indicators applied to paired measurements', *American Journal of Human Biology*, (1994).
7. Bartko, J. J. 'The intraclass correlation coefficient as a measure of reliability', *Psychological Reports*, **19**, 3–11 (1966).
8. Bartko, J. J. 'Corrective note to the intraclass correlation coefficient as a measure of reliability', *Psychological Reports*, **34**, 418 (1974).
9. Shrout, P. E. and Fleiss, J. L. 'Intraclass correlations: Uses in assessing rater reliability', *Psychological Bulletin*, **86**, 420–428 (1979).
10. Snedecor, G. W. and Cochran, W. G. *Statistical Methods*, Sixth Edition, Iowa State University Press, Ames, Iowa, 1976, p. 197.
11. Altman, D. G. and Bland, J. M. 'Measurement in medicine: The analysis of method comparison studies', *The Statistician*, **32**, 307–317 (1983).
12. Altman, D. G. 'Plotting probability ellipses', *Applied Statistics*, **27**, 347–348 (1978).
13. Lin, L. I. ' A concordance correlation coefficient to evaluate reproductibility', *Biometrics*, **45**, 255–268 (1989).