A NOTE ON THE INTERPRETATION OF WEIGHTED KAPPA AND ITS RELATIONS TO OTHER RATER AGREEMENT STATISTICS FOR METRIC SCALES

CHRISTOF SCHUSTER
University of Notre Dame

This article presents a formula for weighted kappa in terms of rater means, rater variances, and the rater covariance that is particularly helpful in emphasizing that weighted kappa is an absolute agreement measure in the sense that it is sensitive to differences in rater's marginal distributions. Specifically, rater mean differences will decrease the value of weighted kappa relative to the value of the intraclass correlation that ignores mean differences. In addition, if rater variances also differ, then the value of weighted kappa will be decreased relative to the value of the product-moment correlation. Equality constraints on the rater means and variances are given to illustrate the relationships between weighted kappa, the intraclass correlation, and the product-moment correlation. In addition, the expression for weighted kappa shows that weighted kappa belongs to the Zegers–ten Berge family of chance-corrected association coefficients. More specifically, weighted kappa is equivalent to the chance-corrected identity coefficient.

Keywords: *weighted kappa; intraclass correlation; rater agreement; association coefficients*

If two raters assign the same targets to categories, the ratings can be arranged in a bivariate frequency table such as Table 1, where for the sake of concreteness three response categories have been assumed. In case categories are ordered along a continuum of values, it is desirable to give partial credit for near agreement. Because weighted kappa (Cohen, 1968) allows for differential weighting of disagreement, it is an attractive agreement statistic for ordered categories and preferable to Cohen's kappa (1960), which distinguishes only between agreement and disagreement cases.

Generally, the gravity of a disagreement is related to the number of categories by which raters differ. One way to implement a weighting scheme that

Educational and Psychological Measurement, Vol. 64 No. 2, April 2004 243-253
DOI: 10.1177/0013164403260197
© 2004 Sage Publications

Table 1
Bivariate Frequency Table of Ratings Obtained From Two Raters

	Rater 2		
	Positive	Neutral	Negative
Rater 1			
Positive	2	1	0
Neutral	0	1	1
Negative	0	1	3

reflects this simple idea is to assign successive integers to the ordered categories. The severity of a disagreement is then determined from the score difference. Popular choices for so-called disagreement weights are the square and the absolute value of the score differences (Agresti, 2002, p. 435). Weighted kappa's emphasis on score differences makes it sensitive to rater mean differences in the sense that these mean differences will decrease the value of weighted kappa relative to the value of the intraclass correlation that ignores rater mean differences. In addition, if rater variances also differ, then the values of weighted kappa and the intraclass correlation will be decreased relative to the value of the product-moment correlation.

To illustrate the sensitivity of weighted kappa to the rater's marginal distributions, consider two equally qualified therapists classifying psychiatric patients at the onset of therapy with respect to expected therapeutic success. Specifically, suppose raters classify patients into one of the following categories: positive, neutral, and negative. If the first therapist tends to give optimistic ratings whereas the second tends to give pessimistic ratings, then a significant proportion of patients could be classified quite differently depending on which therapist rates the patient. Nevertheless, if rater agreement is calculated from a statistic which is insensitive to rater mean differences, such as the product-moment correlation, then the agreement could still attain a very high value. However, because weighted kappa is not only sensitive to rater covariation but also to rater mean differences, its value would be decreased considerably relative to the product-moment correlation. In this sense, weighted kappa is an absolute agreement measure.

As soon as scores are assigned to categories, rater agreement can also be evaluated by arranging the scores in a two-way array such that the rows and columns correspond to targets and raters, respectively, and then partition the rating variability based on a two-way ANOVA model (Ebel, 1951; Guilford, 1954; Shrout & Fleiss, 1979). Specifically, if the jth rater, $j = 1, 2$, assigns the ith target, $i = 1, \ldots, n$, to the cth category, $c = 1, \ldots, r$, then the target's score is denoted $x_{ij}(c)$. Because there are no within-cell replications, the additive ANOVA model

$$x_{ij}(c) = \mu + \tau_i + \theta_j + e_{ij} \quad (1)$$

is selected, where the τ_i represent random target effects and the θ_j represent the rater effects, which may either be considered fixed or random. To illustrate this alternative representation of the ratings, consider the data in Table 1. If the scores for the positive, neutral, and negative categories are denoted as $x(1)$, $x(2)$, and $x(3)$, respectively, then the ratings can be arranged in the two-way array given in Table 2. To be consistent with Equation 1, the scores in Table 2 have a subscript indicating the rater, although this is not strictly necessary because the category scores do not vary across raters, that is, $x_{i1}(c) = x_{i2}(c)$ for all targets and all categories.

Based on the various mean squares of the ANOVA model, one can calculate intraclass correlation coefficients (ICC) to index rater agreement. This approach has two attractive features. First, whereas weighted kappa has been defined by Cohen (1968) for two raters only, intraclass correlations can be calculated for two or more raters. Second, one can define several intraclass correlations that reflect different rater characteristics by changing the assumptions underlying the ANOVA model (McGraw & Wong, 1996). In the present context, two different intraclass correlations result from considering the rater effects either as fixed or random. These two intraclass correlations will be denoted as R and R', respectively. The labels of the two intraclass correlations used in this article are related in the following way to the ICC labels used by Shrout and Fleiss (1979): R is identical to ICC(3, 1), whereas R' is identical to ICC(2, 1).

Because in the case of two raters weighted kappa and intraclass correlations can be calculated from the same data, the question arises as to how these coefficients are related. This issue was first addressed by Fleiss and Cohen (1973), who showed that if the number of targets is large, weighted kappa based on squared disagreement weights is essentially equivalent to the intraclass correlation obtained from considering rater effects random, in our notation R'.

The present article gives a simple expression of weighted kappa (see Equation 5) that involves only the rater means, the rater variances, and the rater covariance. Although the sensitivity of weighted kappa to rater mean differences is well known (e.g., Becker, 2000), the expression of weighted kappa given here exposes this property in a direct way and also makes the role of rater variance differences explicit.

Weighted Kappa

Assume that two raters assign each of n targets to one of r different categories. Let n_{ij} denote the frequency with which the first and second rater assigned targets to categories i and j, respectively. Let $n_{i\bullet}$ denote the total

Table 2
Two-Way Array Representation of the Frequencies Given in Table 1

Target	Rater 1	Rater 2
1	$x_{11}(1)$	$x_{12}(1)$
2	$x_{21}(1)$	$x_{22}(1)$
3	$x_{31}(1)$	$x_{32}(2)$
4	$x_{41}(2)$	$x_{42}(2)$
5	$x_{51}(2)$	$x_{52}(3)$
6	$x_{61}(3)$	$x_{62}(2)$
7	$x_{71}(3)$	$x_{72}(3)$
8	$x_{81}(3)$	$x_{82}(3)$
9	$x_{91}(3)$	$x_{92}(3)$

Note. $x(1)$, $x(2)$, and $x(3)$ denote the scores assigned to the positive, neutral, and negative categories, respectively.

number of targets assigned to the ith category by the first rater, and let $n_{\cdot j}$ be defined similarly for the second rater. The observed proportions n_{ij}/n, $n_{i\cdot}/n$, and $n_{\cdot j}/n$ will be denoted as p_{ij}, $p_{i\cdot}$, and $p_{\cdot j}$, respectively. Finally, let υ_{ij} denote the disagreement weights. Typically, $\upsilon_{ij} = 0$ is selected for cells for which raters agree and $\upsilon_{ij} > 0$ if $i \neq j$, that is, raters show disagreement. Cohen (1968) defined weighted kappa as

$$\kappa_w = 1 - \frac{q_0}{q_e}, \tag{2}$$

where

$$q_0 = \sum_i \sum_j \upsilon_{ij} p_{ij} \tag{3}$$

$$q_e = \sum_i \sum_j \upsilon_{ij} p_{i\cdot} \cdot p_{\cdot j}. \tag{4}$$

Note that q_o can be interpreted as average disagreement and q_e can be interpreted as average disagreement if both raters randomly assign targets in accordance with their respective base rates. Froman and Llabre (1985) have demonstrated the equivalence of weighted kappa and the "del" measure of prediction analysis (Hildebrand, Laing, & Rosenthal, 1977).

One popular option for selecting disagreement weights is to assign integer scores to the categories and then use the squared difference between the row and column scores as the weight for the corresponding cell. More specifically, squared disagreement weights, which will be considered throughout the remainder of this article, are calculated as $\upsilon_{ij} = (i - j)^2$. For the frequencies given in Table 1, Equations 3 and 4 yield the following:

$$q_0 = (1-1)^2\,\tfrac{2}{9} + (1-2)^2\,\tfrac{1}{9} + \cdots + (3-2)^2\,\tfrac{1}{9} + (3-3)^2\,\tfrac{3}{9} = \tfrac{1}{3},$$

$$q_e = (1-1)^2\,\tfrac{6}{81} + (1-2)^2\,\tfrac{9}{81} + \cdots + (3-2)^2\,\tfrac{12}{81} + (3-3)^2\,\tfrac{16}{81} = \tfrac{113}{81},$$

and therefore, $\kappa_\omega = 1 - (1/3)/(113/81) = .761$.

Another Expression for Weighted Kappa

Important properties of weighted kappa are exposed by expressing it in terms of rater means, variances, and the rater covariance. If the data are arranged as in the two-way ANOVA situation (see Table 2), weighted kappa can be represented as the following:

$$\kappa_w = \frac{s_{12}}{\frac{1}{2}(s_1^2 + s_2^2) + \frac{1}{2}\left(\frac{n}{n-1}\right)(\bar{x}_{\cdot 1} - \bar{x}_{\cdot 2})^2}, \tag{5}$$

where $\bar{x}_{\cdot j} = \sum_i x_{ij}(c)$ denotes the mean score of jth rater, j = 1, 2; $s_j^2 = (n-1)^{-1}\sum_i (x_{ij}(c) - \bar{x}_{\cdot 1})^2$ denotes the score variance of the jth rater, j = 1, 2; and $s_{12} = (n-1)^{-1}\sum_1 (x_{i1}(c) - \bar{x}_{\cdot 1})(x_{i2}(c) - \bar{x}_{\cdot 2})$ denotes the score covariance between the two raters. Note that the variance and the covariance formulas are unbiased estimators of the corresponding population parameters.

Equation 5 can be obtained via the following arguments. First, Fleiss and Cohen (1973, p. 616) have shown that for two raters, weighted kappa can be expressed in terms of mean squares obtained from the ANOVA model of Equation 1. Specifically, they gave the expression

$$\kappa_w = \frac{\text{TMS} - \text{EMS}}{\text{TMS} + \text{EMS} + \left(\frac{2}{n-1}\right)\text{RMS}}, \tag{6}$$

where TMS, RMS, and EMS refer to the target-, rater-, and error mean squares, respectively, which are based on $(n-1)$, $(k-1)$, and $(n-1)(k-1)$ degrees of freedom, respectively. Thus, it is sufficient to show that the right sides of Equations 5 and 6 are identical. It will be convenient to prove the equality in three steps:

$$\begin{aligned} &(i) \quad \left(\tfrac{2}{n-1}\right)\text{RMS} = \tfrac{n}{n-1}(x_{\cdot 1} - x_{\cdot 2})^2, \\ &(ii) \quad \text{TMS} + \text{EMS} = s_1^2 + s_2^2, \\ &(iii) \quad \text{TMS} - \text{EMS} = 2s_{12}. \end{aligned} \tag{7}$$

Of course, if all three steps hold, then the equivalence of Equations 5 and 6 follows immediately.

To see that (*i*) holds simply note that the rater mean squares for two raters is RMS = $\left[(\bar{x}_{.1} - \bar{x}_{..})^2 + (\bar{x}_{.2} - \bar{x}_{..})^2\right]$, where $\bar{x}_{..}$ denotes the grand mean. Simple algebraic manipulations yield the expression RMS = RMS = (n/2) $(\bar{x}_{.1} - \bar{x}_{.2})^2$, and the result given in (*i*) of Equation 7 follows.

To show (*i*) and (*ii*), it is natural to consider a table having k columns and obtain these steps for the special case of $k = 2$. Both conditions follow from two results given in Winer (1971, p. 291). These are the following:

$$\text{TMS} = \overline{\text{var}} + (k - 1)\overline{\text{cov}}, \tag{8}$$

$$\text{EMS} = \overline{\text{var}} - \overline{\text{cov}}, \tag{9}$$

where $\overline{\text{var}}$ is the average rater variance and $\overline{\text{cov}}$ the average rater covariance where the average is taken over all possible rater pairs. Calculating the sum of TMS and EMS using Equations 8 and 9 yields TMS + EMS = $2\overline{\text{var}} + (k-2)$. If $k = 2$, then (*ii*) follows immediately. Similarly, calculating the difference between TMS and EMS yields TMS – EMS = $k\overline{\text{cov}}$. Clearly, if $k = 2$, this expression is equal to (*iii*).

Equation 5 can also be used to calculate weighted kappa from the array of scores given in Table 2, where the scores are obtained from assigning successive integers to the categories. Thus, $x_{ij}(1) = 1$, $x_{ij}(2) = 2$, and $x_{ij}(3) = 3$. The assignment of integer scores can be expressed more concisely as $x_{ij}(c) = c$ for all i and j. Based on these scores, one obtains for the means $\bar{x}_{.1} = 2.111$ and $\bar{x}_{.2} = 2.222$, for the variances $s_1^2 = 0.861$ and $s_2^2 = 0.694$, and for the covariance $s_{12} = 0.597$. Inserting these values into Equation 5 confirms $\kappa_\omega = .761$.

Relations Among κ_ω, R, and r

Consider two special cases of Equation 5 that emerge from equality constraints. First, assuming equal rater means yields

$$\kappa_w^{(1)} = \frac{s_{12}}{\frac{1}{2}(s_1^2 + s_2^2)}, \tag{10}$$

where the superscript is used to indicate that this equation holds only if rater means are equal. The equivalence of this expression to R, which in case of two raters is defined as (e.g., Shrout & Fleiss, 1979)

$$R = (\text{TMS} - \text{EMS})/(\text{TMS} + \text{EMS}), \tag{11}$$

can readily be verified by noting that because equal rater means imply RMS = 0, Equation 6 simplifies to Equation 11. Thus, equal rater means imply $\kappa_w^{(1)} = \text{P}$.

Second, assume that both the rater means and the rater variances are equal. Cohen (1968, p. 218) noted that these conditions imply the equivalence of weighted kappa and the product-moment correlation, *r*, calculated from integer scores. To see this, note that by equating the means and variances in Equation 5, one obtains

$$\kappa_w^{(2)} = s_{12} / s^2, \tag{12}$$

where s^2 denotes the common variance and the superscript is used to indicate that this equation holds only under equal rater means and equal rater variances. Alternatively, one can arrive at this expression by equating the variances in the usual formula of the product-moment correlation, which is $r = s_{12} / \sqrt{s_1^2 s_2^2}$. Thus, these conditions imply $\kappa_w^{(2)} = r$.

In addition, the comparison of the intraclass correlation *R*, which is equivalent to $\kappa_w^{(2)} = r$, and the product moment correlation implies that the product moment correlation establishes the upper limit of weighted kappa. This follows from noting that the denominator of Equation 10 can be considered the arithmetic mean of the rater variances, whereas the denominator of the product moment correlation, $r = s_{12} / \sqrt{s_1^2 s_2^2}$, can be considered the geometric mean of the rater variances. Because the geometric mean is always lower or equal to the arithmetic mean, it follows that unequal rater variances will decrease weighted kappa relative to the value of the product moment correlation.

One can also compare weighted kappa, the intraclass correlation that ignores rater mean differences, and the product-moment correlation in terms of score transformations that will not affect their value. First, it is well known that the product-moment correlation is invariant to linear transformations of the variables. More specifically, if $y_{i1} = a_1 x_{i1} + b_1$ and $y_{i2} = a_2 x_{i2} + b_2$—where we assume $a_1 > 0$ and $a_2 > 0$ to avoid pathological cases—then the correlation between the *y*-scores is the same as the correlation between the original *x*-scores. Note that if a_1 and a_2 as well as b_1 and b_2 are allowed to differ, then the transformations can always achieve equal means and variances of the *y*-scores. This confirms that the product moment correlation is insensitive to differences in rater means and variances.

Second, consider the score transformations that maintain the value of the intraclass correlation *R*. From Equation 10, which has just been shown to be equivalent to *R*, it is easy to see that linear score transformations do not change the value of *R* provided $a_1 = a_2$. Note that if only b_1 and b_2 are allowed to differ, then the transformations can always achieve equal means of the *y*-scores, but not equal variances. This confirms that the intraclass correlation *R* is insensitive to rater mean differences but sensitive to rater variance differences. Third, it can easily be verified from the defining equation of weighted kappa (see Equation 2) or from Equation 5 that linear score trans-

Table 3
Sensitivity of r, R, *and* κ_ω *to Linear Score Transformations*

	Sensitive to Changes in	
Coefficient	Scale	Location
r	No	No
R	Yes	No
κ_ω	Yes	Yes

formations do not alter the value of weighted kappa, provided that both $a_1 = a_2$ and $b_1 = b_2$. In other words, only if the same linear transformation is applied to the *x*-scores is the value of weighted kappa unaffected. Thus, weighted kappa is sensitive to mean and variance differences between raters. Table 3 summarizes the sensitivity properties of the three coefficients to linear score transformations.

Finally, the following heuristic argument that is based on the value preserving score transformations can be used to confirm the ordering relations between κ_ω, *R*, and *r* outlined above. Because one may consider each of the coefficients as measuring the degree to which linear score transformations are able to produce a close match between the *y*-scores, the increase in the restrictiveness of the permissible transformations as one moves from *r* to *R* to κ_ω leads to a decrease in their respective values such that *r* is at least as high as *R*, which is at least as high as κ_ω.

Weighted Kappa in the Context of the Zegers–ten Berge Family of Association Coefficients

Zegers (1986) presented chance-corrected versions of association coefficients suggested by Zegers and ten Berge (1985). This theory presents a general formula for association coefficients for metric scales and derives four special cases that correspond to different scale levels. First, absolute scales do not allow any score transformation (other than the trivial identity transformation); second, additive scales allow the addition of a constant to all scores; third, ratio scales allow the multiplication of all scores with a positive constant; and fourth, interval scales allow linear scale transformations. For absolute, additive, ratio, and interval scales, the coefficients of identity, additivity, proportionality, and linearity are defined, respectively. In the context of rater agreement data, the Zegers–ten Berge theory has been applied by Stine (1989) and Zegers (1991).

It has been shown that numerous association coefficients for metric scales are members of this family of coefficients. In fact, *R* and *r* are identical to the

coefficient of additivity and linearity, respectively. It will now be shown that weighted kappa also belongs to the Zegers–ten Berge family of association coefficients. To see this, consider the biased variance estimator that is obtained by dividing the sum of squared deviations by the sample size n instead of $n - 1$. Similarly, the biased covariance estimator will be considered that is obtained by dividing the sum of deviation products by the sample size n instead of $n - 1$. If the biased versions of the variance and the covariance are denoted as $\tilde{s}_j^2, j = 1, 2$, and $\tilde{s}_{12}$, then the unbiased and biased versions are simply related by $\tilde{s}_1^2 = s_1^2(n-1)/n$, $\tilde{s}_2^2 = s_2^2(n-1)/n$, , and $\tilde{s}_{12} = s_{12}(n-1)/n$. Replacing the unbiased estimators with the biased estimators in Equation 5 yields the following:

$$\kappa_w = \frac{\tilde{s}_{12}}{\frac{1}{2}\left(\tilde{s}_1^2 + \tilde{s}_2^2\right) + \frac{1}{2}(\bar{x}_{.1} - \bar{x}_{.2})^2}, \tag{13}$$

which is almost identical to Equation 5. In fact, Equation 13 is conceptually simpler because it does not involve the factor $n/(n-1)$ in the denominator of Equation 5.

This alternative expression for weighted kappa shows that weighted kappa belongs to the Zegers–ten Berge family of association coefficients because it is identical to the chance-corrected identity coefficient of Zegers (1986, p. 560, Equation 6). In addition, Equation 13 is identical to the coefficient of equality proposed by Jobson (1976, p. 272, Equation 2).

Whereas weighted kappa was originally defined for scales having only a relatively small number of categories, the Zegers–ten Berge theory applies equally well to continuous scales. However, one can also define the average disagreement and average chance disagreement (see Equations 3 and 4) based on the squared distance between the rater scores more generally as

$$q_0 = E(X - Y)^2 \tag{14}$$

$$q_e = E_c(X - Y)^2, \tag{15}$$

where E_c indicates that the expectation should be taken assuming X and Y are independent in their joint distribution. Rewriting q_0 as

$$q_0 = E\{(X - \mu_1) - (Y - \mu_2) + (\mu_1 - \mu_2)\}^2, \tag{16}$$

where μ_1 and μ_2 denote the population means of the X and Y scores, respectively, yields after some algebra:

$$q_0 = \sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2 - 2\sigma_{12}, \tag{17}$$

where σ_1^2, σ_2^2, and σ_{12} denote the population variances of the *X* and *Y* scores and the population covariance, respectively. Because q_e can be obtained as the special case of q_0 for which $\sigma_{12} = 0$, the defining equation of weighted kappa (see Equation 2) yields the following:

$$\kappa_w = 1 - \frac{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2 - 2\mu_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \tag{18}$$

$$= \frac{\sigma_{12}}{\frac{1}{2}(\sigma_1^2 + \sigma_2^2) + \frac{1}{2}(\mu_1 - \mu_2)^2} \tag{19}$$

which is identical to Equation 13 except that the sample estimates have been replaced with the corresponding population quantities.

Practical Considerations

It is important that researchers are aware that by selecting among popular rater agreement statistics discussed in this article, they decide on whether mean differences and/or variance differences in ratings should be ignored or accounted for. Stine (1989) as well as Zegers (1991) have commented on the rationale for selecting one statistic over another.

Stine (1989) has argued that, in general, association coefficients should be based only on meaningful relations between ratings. Therefore, if ratings lie on an interval scale, then because the origin and the scale unit are arbitrary the proper agreement statistic is *r*. If the ratings lie on an additive scale, then because the origin is arbitrary the proper agreement statistic is *R*. Finally, if the ratings lie on an absolute scale, then because neither the scale origin nor the scale unit is arbitrary the agreement statistic should be sensitive to rater differences in the scale location as well as rater differences in the scale units. Because weighted kappa fulfills these requirements, it is the proper agreement statistic for absolute scales.

However, because it is impossible to determine the scale level of ratings, Zegers (1991) advocated a pragmatic approach in which the researcher decides which information is meaningful and which is irrelevant. For instance, if a pass/fail decision has to be reached by comparing a score against a certain cutoff value, then rater mean differences are important. Thus, a rater agreement statistic should reflect such rater response tendencies by decreasing its value relative to agreement coefficients that ignore response tendencies. However, if a graded performance reward should be given to individuals, then perfect agreement between raters would be reached if raters agreed only on the performance ordering of the individuals. Thus, only the relative location of individuals within each of the two orderings would be important; therefore, the assessment of the rater agreement should ignore response tendencies between raters.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Becker, G. (2000). Creating comparability among reliability coefficients: The case of Cronbach alpha and Cohen kappa. *Psychological Reports*, *87*, 1171-1182.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213-220.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika*, *16*(4), 407-424.

Froman, T. W., & Llabre, M. M. (1985). The equivalence of kappa and del. *Perceptual and Motor Skills*, *60*, 3-9.

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

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. (1977). *Prediction analysis of cross-classifications*. New York: Wiley.

Jobson, J. D. (1976). A coefficient for questionnaire items with interval scales. *Educational and Psychological Measurement*, *36*, 271-274.

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428.

Stine, W. W. (1989). Interobserver relational agreement. *Psychological Bulletin*, *106*(2), 341-347.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Zegers, F. E. (1986). A family of chance-corrected association coefficients for metric scales. *Psychometrika*, *51*(4), 559-562.

Zegers, F. E. (1991). Coefficients for interrater agreement. *Applied Psychological Measurement*, *15*(4), 321-333.

Zegers, F. E., & ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika*, *50*(1), 17-24.