PSYCHOMETRIKA—VOL. 58, NO. 2, 357–370
JUNE 1993

A GENERALIZED FAMILY OF COEFFICIENTS OF RELATIONAL AGREEMENT FOR NUMERICAL SCALES

ROBERT F. FAGOT

UNIVERSITY OF OREGON

A family of coefficients of relational agreement for numerical scales is proposed. The theory is a generalization to multiple judges of the Zegers and ten Berge theory of association coefficients for two variables and is based on the premise that the choice of a coefficient depends on the scale type of the variables, defined by the class of admissible transformations. Coefficients of relational agreement that denote agreement with respect to empirically meaningful relationships are derived for absolute, ratio, interval, and additive scales. The proposed theory is compared to intraclass correlation, and it is shown that the coefficient of additivity is identical to one measure of intraclass correlation.

Key words: agreement coefficient, meaningfulness theory, scale type, intraclass correlation.

Zegers and ten Berge (1985) proposed a general formula for bivariate association coefficients for metric scales, and Zegers (1986) proposed a chance-corrected version of the general formula. Their theory was based on the premise that the choice of an association coefficient between two variables depends on the scale type of the variables, defined by the class of admissible transfomations. Stine (1989a) extended the Zegers-ten Berge theory to several additional scales not considered by Zegers and ten Berge, and shifted the focus from "association" between two variables to "relational agreement" between observers. Stine (1989a) also demonstrated the utility of concepts of *meaningfulness* in assessing interobserver agreement.

The Zegers-ten Berge theory as well as the extended theory of Stine (1989a) deals with association or agreement between only two variables or observers. This paper proposes a generalization of the Zegers-ten Berge theory to the case of multiple judges, with meaningfulness theory playing a central role. The typical application will be to data matrices in which two or more observers or judges make numerical ratings on each of a number of "targets." The proposed theory will be examined in relation to *intraclass correlation*, which also provides a theory and procedure for assessing interobserver agreement for many observers (Fagot, 1991; Shrout & Fleiss, 1979).

Scale Type, Admissible Transformations, and Relational Agreement

Table 1 lists the four metric scales considered by Zegers and ten Berge (1985) together with the admissible (defining) transformation for each (see Stine, 1989a, Table 1, for other possible scales with defining transformations). For example, the ratio scale permits only multiplication by a positive constant (change of unit); the interval scale permits positive linear (affine) transformations (change of unit and zero point); the additive scale permits only translation (additive constant); and the absolute scale permits no transformation of the original scale. (We note that the name "additive" scale was used by Zegers and ten Berge, 1985; and followed by Stine, 1989a; but was earlier

The author thanks the Editor and anonymous reviewers for helpful suggestions. Request for reprints should be sent to Robert F. Fagot, Department of Psychology, University of Oregon, Eugene, OR 97403-1227.

0033-3123/93/0600-92434$00.75/0

TABLE 1

Scales

Scale	Admissible (Defining) Transformation
Absolute	$\varphi' = \varphi$
Additive	$\varphi' = \varphi + \alpha$
Ratio	$\varphi' = \beta\varphi$
Interval	$\varphi' = \beta\varphi + \alpha$

Note. φ and φ' are representing functions;
α is a real number and β is a positive real number.

called the "difference" scale by Suppes and Zinnes, 1963. It will be shown later that there is a formal link between the ratio and additive scales.)

The basis for the central role of scale type in constructing a coefficient of interobserver agreement is precisely this difference among scales in admissible tranforms, and therefore, in the number and kind of arbitrary scale factors. For example, suppose the set of ordered pairs (1,16), (2,32), (3,48) denotes the ratings of two judges on three targets. There appears to be very poor agreement between the judges since the ratings for each target do not match. But if the ratings are for the weight of objects with the first judge rating in pounds and the second in ounces, then the judges are in perfect agreement relative to a ratio scale. In other words, according to this theory, a sufficient condition for perfect agreement on a ratio scale is the *proportionality* of the judges' ratings. In general, perfect agreement signifies that the judges' ratings are related by an admissible transformation relative to scale type, and the coefficient of agreement should estimate the degree to which this relation holds.

These ideas are embodied in Stine's concept of *relational agreement* (Stine, 1989a) that denotes agreement with respect to *empirically meaningful relationships*. In any application there will be both meaningful and meaningless disagreement. Thus, in the previous example, if ratings are made on a ratio scale, then proportionality of the judges' ratings is meaningless disagreement (reflecting differences in units permitted by a ratio scale), but additive disagreement is meaningful disagreement (since the origin is

fixed for a ratio scale). The idea is that a coefficient of relational agreement should vary with (be attenuated by) meaningful disagreement but be independent of meaningless disagreement. (Modern ideas about meaningful relations owe a great debt, of course, to Stevens, 1946, 1951, 1968; Stine, 1989b, provides a good introduction to the subject of meaningful inference—also, see Adams, Fagot, & Robinson, 1965; Luce, Krantz, Suppes, & Tversky, 1990, chap. 22; Narens, 1985; Roberts, 1979; and Suppes & Zinnes, 1963).

Zegers-ten Berge Theory

Zegers and ten Berge (1985) proposed a general formula for association coefficients between two variables for the metric scales listed in Table 1. They required that the association coefficient be invariant under admissible transformations of the scale and sensitive to nonadmissible transformations. Key to the development of the general formula was the concept of a *uniformed version* of the variables that would satisfy the conditions of invariance under admissible transformations and sensitivity to nonadmissible transformations. A *uniforming transformation* is a member of the class of admissible transformations for a given scale type.

Let X_i denote a rating on the i-th variable before transformation and V_i the uniformed version. Then the uniforming transformations are:

$$V_i = X_i, \qquad \text{(absolute scale)} \qquad (1a)$$

$$V_i = X_i - \bar{X}_i, \qquad \text{(additive scale)} \qquad (1b)$$

$$V_i = X_i/T_i, \qquad \text{(ratio scale)} \qquad (1c)$$

$$V_i = (X_i - \bar{X}_i)/S_i, \qquad \text{(interval scale)} \qquad (1d)$$

where $\bar{X}_i$ and S_i are the arithmetic mean and standard deviation of variable (judge) i, and $T_i = [(1/n)\ \Sigma_{t=1}^{n} X_{it}^2]^{1/2}$, where n is the number of ratings on each target t. With slight change in notation, (1a), (1b), (1c), and (1d) correspond to (2a), (2b), (2c), and (2d), respectively, in Zegers and ten Berge (1985). The additive transform (1b) centers the variable at zero, the multiplicative transformation (1c) rescales the variable to obtain a mean squared value of one, and the linear transform (1d) is the familiar "standardizing" Z transform (with zero mean and standard deviation of one).

The idea behind the uniforming transformations is that after the appropriate uniforming transform is applied to the X_i, then if the uniformed versions should match, there is perfect agreement among the judges; and the coefficient of association estimates the degree to which the uniformed versions agree (in numerical value). For example, if the set of ordered pairs (1,2), (2,4), and (3,6) denotes the (proportional) ratings of two judges for three targets on a ratio scale, then applying transform (1c), $T_1 = 2.16$, $T_2 = 4.32$, and the uniformed versions are (.463,.463), (.926,.926), and (1.389,1.389), indicating perfect agreement on a ratio scale. Although Zegers and ten Berge do not state their results formally in terms of meaningfulness theory, it is clear that the uniformed versions have the effect of insuring that the coefficients vary with meaningful disagreement but are independent of meaningless disagreement (relative to scale type).

Based on a mean squared difference function of the uniformed versions, Zegers and ten Berge derived their general formula of association coefficients for metric scales (g_{ij}) between two uniformed versions V_i and V_j (their Equation (8) with slight changes in notation):

$$g_{ij} = \frac{2\sum_{t=1}^{n} V_{it}V_{jt}}{\sum_{t=1}^{n} V_{it}^2 + \sum_{t=1}^{n} V_{jt}^2}. \tag{2}$$

Zegers (1986) then derived a general formula g'_{ij} (his Equation (5)) for the bivariate case, correcting (2) for chance agreement (see Fagot & Mazo, 1989). By substituting the appropriate uniforming transformation in the formula for g_{ij} and g'_{ij}, association coefficients were obtained for each of the scales listed in Table 1.

For the bivariate case, the coefficients of Additivity and Linearity (Pearson r) did not require correction for chance agreement (Zegers & ten Berge, 1985, Equations (19) and (21), respectively). The coefficients of Identity (for absolute scales) and Proportionality (for ratio scales) corrected for chance agreement are given in Zegers (1986, Equations (6) and (7), respectively).

Generalization of Zegers-ten Berge Theory

The derivation of formulas for agreement coefficients for the case $k > 2$ involves four steps: (a) the generalization of (2) to $g(k)$, the general formula; (b) finding $g_c(k)$, the value of $g(k)$ for chance agreement; (c) deriving the *family* of agreement coefficients, corrected for chance agreement, from (a) and (b); and (d) derivation of the special formulas for each scale in Table 1, based on the family (general formula) of corrected coefficients of agreement. I begin with a note on notation.

Let X_{it} denote the rating of target t by judge i, and V_{it} the corresponding uniformed version. Summation throughout will be $t = 1, \ldots, n$ and $i, j = 1, \ldots, k$. The summation $\Sigma_{i<j}^{k}$ denotes summation over the $(k/2)(k-1)$ i,j pairs. For example if $k = 3$, then summation is over the three i,j pairs (1,2), (1,3), and (2,3) and $\Sigma_{i<j}^{k}\Sigma_{t=1}^{n} X_{it}X_{jt} = \Sigma_{t=1}^{n} X_{1t}X_{2t} + \Sigma_{t=1}^{n} X_{1t}X_{3t} + \Sigma_{t=1}^{n} X_{2t}X_{3t}$.

A generalization of (2) is

$$g(k) = \frac{2c\sum_{i<j}^{k}\sum_{t=1}^{n} V_{it}V_{jt}}{\sum_{i=1}^{k}\sum_{t=1}^{n} V_{it}^2}, \tag{3}$$

where c is a constant to be determined by appropriate constraint. The constant c is uniquely determined by constraining $g(k) = 1$ if $V_{it} = V_{jt}$, for all i, j, t (i.e., if for each target the uniformed versions match in numerical value). Then given this constraint, the numerator of (3) is $2c(k/2)(k-1)\,\Sigma_t V_{it}^2$, the denominator of (3) is $k\,\Sigma_t V_{it}^2$, and hence, $c = 1/(k-1)$. Substituting this value of c in (3), one gets

$$g(k) = \frac{2\sum_{i<j}^{k}\sum_{t=1}^{n} V_{it}V_{jt}}{(k-1)\sum_{i=1}^{k}\sum_{t=1}^{n} V_{it}^2}, \tag{4}$$

which is the general formula for agreement coefficients for multiple judges ($k \geq 2$). However, this formula capitalizes on chance for some members of the family of coefficients, and hence, $g(k)$ must be corrected for chance agreement.

A well-known procedure for chance correction of an agreement coefficient, expressed in terms of $g(k)$ is

$$g'(k) = \frac{g(k) - g_c(k)}{1 - g_c(k)}, \tag{5}$$

where $g'(k)$ is the chance-corrected coefficient and $g_c(k)$ represents "chance agreement". The numerator of (5) may be interpreted as representing agreement in excess of what is to be expected by chance, and the denominator as the maximum possible excess. Hence, (5) represents the proportional excess beyond what is expected by chance.

This form of correction has been used with Cohen's kappa (Cohen, 1960); Guttman's coefficient of reproduction (Green, 1956); for correction in the overestimation of the multiple correlation coefficient (Fisher, 1924); and more importantly for our purposes, by Zegers (1986) in correcting the bivariate general formula of association coefficients in (2) for chance agreement.

Zegers (1986) treated chance expectation g_c in the bivariate case as the expectation of g_{ij} in (2) over all $n!$ permutations of the scores on one variable for a given sample, with the other variable fixed (Kendall & Stuart, 1961, p. 474). Then for the bivariate case, treating V_i as fixed and taking all possible permutations over V_j, the expectation of the numerator is $E[2\ \Sigma_t\ V_{it}V_{jt}] = 2\ \Sigma_t\ V_{it}\bar{V}_j = 2n\ \bar{V}_i\bar{V}_j$, where $\bar{V}_i$ and $\bar{V}_j$ are arithmetic means of the uniformed versions for i and j. The denominator does not change over permutations of V_j.

Extending the approach to the multivariate case, one gets $E[2\ \Sigma_{i<j}^{k}\ \Sigma_t^n\ V_{it}V_{jt}] = 2n\ \Sigma_{i<j}^{k}\ \bar{V}_i\bar{V}_j$, and hence, the chance value $g_c(k)$ is

$$g_c(k) = \frac{2n \sum_{i<j}^{k} \bar{V}_i\bar{V}_j}{(k-1) \sum_{i}^{k} \sum_{t}^{n} V_{it}^2}. \tag{6}$$

Substituting $g_c(k)$ from (6) and $g(k)$ from (4) into (5), one gets the general coefficient of agreement corrected for chance:

$$g'(k) = \frac{2\left[\sum_{i<j}^{k} \sum_{t}^{n} V_{it}V_{jt} - n \sum_{i<j}^{k} \bar{V}_i\bar{V}_j\right]}{(k-1) \sum_{i}^{k} \sum_{t}^{n} V_{it}^2 - 2n \sum_{i<j}^{k} \bar{V}_i\bar{V}_j}. \tag{7}$$

Setting $k = 2$, one gets the Zegers general bivariate coefficient as a special case (Zegers, 1986, Equation (5)).

By substituting the appropriate uniforming transformations into (7), the agreement coefficient is obtained for each of the four scale types, using the symbols "I" for identity, "P" for proportionality, "A" for additivity, and "L" for linearity.

Substituting (1a) into (7), one gets the *coefficient of Identity* (I) for absolute scales:

$$I = \frac{2\left[\sum_{i<j}^{k}\sum_{t}^{n} X_{it}X_{jt} - n\sum_{i<j}^{k}\bar{X}_i\bar{X}_j\right]}{(k-1)\sum_{i}^{k}\sum_{t}^{n} X_{it}^2 - 2n\sum_{i<j}^{k}\bar{X}_i\bar{X}_j}, \tag{8a}$$

$$I = \frac{2\sum_{i<j}^{k} S_{ij}}{(k-1)\sum_{i}^{k} S_i^2 + \sum_{i<j}^{k}(\bar{X}_i - \bar{X}_j)^2}, \tag{8b}$$

where $\bar{X}_i$ and S_i^2 denote the sample mean and variance of observer i, and S_{ij} the sample covariance of observers i and j, in all cases of the untransformed scores X.

Substituting (1b) into (7), one gets the *coefficient of Additivity* (A) for additive scales:

$$A = \frac{2\sum_{i<j}^{k}\sum_{t}^{n}(X_{it} - \bar{X}_i)(X_{jt} - \bar{X}_j)}{(k-1)\sum_{i}^{k}\sum_{t}^{n}(X_{it} - \bar{X}_i)^2}, \tag{9a}$$

$$A = \frac{2\sum_{i<j}^{k} S_{ij}}{(k-1)\sum_{i}^{k} S_i^2}. \tag{9b}$$

Substituting (1c) into (7), one gets for the *coefficient of Proportionality* (P) for ratio scales:

$$P = \frac{\sum_{i<j}^{k}\left[\dfrac{\sum_{t}^{n} X_{it}X_{jt} - n\bar{X}_i\bar{X}_j}{\left(\sum_{t}^{n} X_{it}^2 \sum_{t}^{n} X_{jt}^2\right)^{1/2}}\right]}{\left(\frac{k}{2}\right)(k-1) - n\sum_{i<j}^{k}\left[\dfrac{\bar{X}_i\bar{X}_j}{\left(\sum_{t}^{n} X_{it}^2 \sum_{t}^{n} X_{jt}^2\right)^{1/2}}\right]} \tag{10}$$

Substituting (1d) into (7), one gets the coefficient of Linearity (L) for interval scales:

$$L = \frac{2 \sum_{i<j}^{k} \sum_{t}^{n} Z_{it} Z_{jt}}{(k-1) \sum_{i} \sum_{t} Z_{it}^2}, \tag{11a}$$

and since $\Sigma_i \Sigma_t Z_{it}^2 = nk$, and $r_{ij} = \Sigma_t Z_{it} Z_{jt}/n$, $\underline{L} = \Sigma_{i<j}^{k} r_{ij}/[(k/2)(k-1)]$, and

$$L = \bar{r}_{ij}, \tag{11b}$$

the arithmetic mean of the pairwise Pearson r.

It is apparent from inspection of (1b), (1d), and (6) that chance agreement is zero for the coefficients of Additivity and Linearity since in both cases the means of the uniformed versions are zero. This means, of course, that the coefficients of Additivity and Linearity may also be derived by substituting (1b) and (1d), respectively, into (3), the uncorrected general coefficient.

If $k = 2$ in the above formulas for I, A, P, and L, one gets the Zegers-ten Berge bivariate coefficients (designated e', a, p', and r, respectively) as special cases.

Ordering on Coefficients

For the bivariate case, a partial ordering exists among the corrected coefficients: no ordering exists between the corrected coefficients of proportionality and additivity, but each exceeds or equals the coefficient of identity in absolute value. The Pearson r exceeds or equals all three corrected coefficients in absolute value (Zegers, 1986).

For $k > 2$ the only ordering is $|A| \geq |I|$, which is easily seen from inspection of (8b) and (9b). Fagot and Mazo (1989) point out that the ordering of the coefficients should depend on the number of arbitrary scale factors, a rule that works in the bivariate case. But for $k > 2$, the coefficients depend on the pairwise coefficients in a complex way, and in the presence of strong interaction with coefficients negative or near zero, a stable ordering on the coefficients will not occur. However, for high positive values the coefficients tend to have the same partial ordering as in the bivariate case: P and A greater than or equal to I; no ordering on P and A; and L higher than all other coefficients.

Equivalence of Ratio and Additive Scales

A formal connection between the ratio and additive scales should be noted. Narens (1981a, 1981b) and Alper (1985) showed that for the domain of real numbers, the additive scale and the ratio scale are isomorphic (equivalent): taking logs in the ratio case (similarity group) leads to the additive (difference) transformations (translation group). Thus the uniforming transformation for the additive scale (1b) is equivalent to $V_i = X_i - \log T_i$, where T is the parameter from (1c) for the ratio scale. However, this formal link need not be interpreted to imply that the additive scale is empirically redundant. Clearly, the coefficient of relational agreement will differ for the two scales. Furthermore, there will be situations for which the additive scale is more natural; for example, in studies requiring subjects to give direct estimates of stimulus *differences*. In fact, choosing between ratio and difference *representations*, based on ratio and difference judgments, is an important ongoing problem in psychological measurement (Birnbaum, 1982).

Alternative Generalization

The generalization $g(k)$ given by (4) was derived from (2) by forming a ratio of sums (a sum of cross products divided by a sum of squares). An alternative form (suggested for consideration by the Editor) is a simple mean of the ratios. The relationship between these two forms needs to be examined.

The generalization for the alternative "mean of ratios" form $g_{MR}(k)$ is:

$$g_{MR}(k) = \left[\frac{4}{k(k-1)}\right] \sum_{i<j}^{k} \left[\frac{\sum_{t}^{n} V_{it}V_{jt}}{\sum_{t}^{n} V_{it}^2 + \sum_{t}^{n} V_{jt}^2}\right].$$

It can be shown that if $\Sigma_t^n V_{it}^2 = \Sigma_t^n V_t^2$ (constant for all i), then $g(k) = g_{MR}(k)$; that is, the two forms give identical generalizations if all judge sums of squares are equal (for the uniformed versions).

The constancy of $\Sigma_t^n V_{it}^2$ is necessarily satisfied for interval and ratio scales. For interval scales, $V_i = Z_i$, $\Sigma_t^n V_{it}^2 = \Sigma_t^n Z_{it}^2 = n$, and hence, for interval scales, the two forms are equivalent. The same conclusion holds for ratio scales: $V_{it} = X_{it}/T_{it}$, $V_{it}^2 = nX_{it}^2/\Sigma_t^n X_{it}^2$, and $\Sigma_t^n V_{it}^2 = (n \Sigma_t^n X_{it}^2)/\Sigma_t^n X_{it}^2 = n$.

Therefore for the two most commonly used numerical scales—ratio and interval—the "ratio of sums" generalization $g(k)$ and the "mean of ratios" generalization $g_{MR}(k)$ are equivalent.

The constancy of $\Sigma_t^n V_{it}^2$ is not, in general, satisfied for the additive scale ($V_i = x_i$) or the absolute scale ($V_i = X_i$), and hence, estimates of A and I will differ for the two forms, although differences will be small for moderate inequalities in $\Sigma_t^n V_{it}^2$.

Further justification for the generalization adopted in this paper (see (4)) is the link to intraclass correlation (developed in the next section), and the analogous definition of strength of effect, in the analysis of variance, as a *ratio of sums* (of squares).

Generalized Agreement Coefficients and Intraclass Correlation

The family of agreement coefficients developed for metric scales has interesting relationships to intraclass correlation. Of particular interest is the identity

$$\text{ICC}(3,1) = A, \tag{12}$$

(see Appendix for ICC(3,1) formula and proof). The two coefficients in (12) are algebraically identical in spite of the fact that they were derived from different assumptions. ICC(3,1) assumes an ANOVA model with judges a fixed factor and targets a random factor, whereas the coefficient of Additivity as a descriptive estimator is neutral with respect to these assumptions. Such assumptions, however, become relevant when statistical inference with respect to A as well as to the other agreement coefficients is introduced. However, one important consequence of (12) is to make explicit the fact that ICC(3,1) is a coefficient of relational agreement only for additive scales. ICC(3,1) is identical to Winer's "adjustment for anchor points" method of estimating ICC (Winer, 1971, pp. 289–293) that explicitly removes the main effect of judges. Note that eliminating the main effect of judges removes only the additive effect of judges, and in the context of the meaningfulness approach, this step is justified only if the additivity

of judges' ratings is treated as meaningless disagreement—as required by the additive scale.

Limits of Coefficients

It is easy to show that the upper limit of all agreement coefficients (I, P, A, L) and ICC(3,1), ICC(2,1) is +1. Lahey, Downey, and Saal (1983) showed that the lower limit of ICC(3,1) is $-1/(k-1)$. If the only source of variability is interaction, then the lower limit of ICC(2,1) is given by $-1/[k-1-(k/n)]$. Then ICC(2,1) < 0 if $k > n/(n-1)$ and ICC(2,1) approaches minus infinity as $n \to -k(1-k)$. But for n large, ICC(2,1) has approximately the lower limit of $-1/(k-1)$.

The lower limit for I, P, A, and L is also $-1/(k-1)$. Of course, since A = ICC(3,1), then obviously the lower limit of A is $-1/(k-1)$, which occurs if the target mean square is zero (BMS = 0; see Appendix).

Inspection of (8b) and (9b) shows that $I = A$ if all judge means are equal. Hence I reaches a lower limit of $-1/(k-1)$ if both main effects are zero (i.e., all variability is error variability).

If the only source of variability is error variance (interaction) and equal judge variability holds, then both P and L have the same lower limit of $-1/(k-1)$. Equal variability for P requires that the $\Sigma_t^n X_{it}^2$ are equal for all i judges; and for L that all judge variances are equal.

Table 2 shows three cases exhibiting lower limits for the coefficients, which are $-1/(k-1) = -.50$ for three judges. For Case I the lower limit for A and ICC(3,1) is reached, but not for the other coefficients because the judge means are not equal. For Case II, the judge means are equal, so in addition $I = -.50$ but still not P and L because judge variability is unequal. For Case III, judge and target means are equal and so is judge variability, so now all coefficients except ICC(2,1) reach the theoretical lower limit of $-.50$. ICC(2,1) $= -1/[k-1-k/n] = -1.00$ for $k = n = 3$. For $n = 10$, $k = 3$, ICC(2,1) $= -.59$.

Comparison of ICC(3,1) and ICC(2,1)

It has been pointed out that ICC(3,1) = A, and hence, ICC(3,1) is a member of the family of coefficients of relational agreement defined by (7). ICC(2,1), on the other hand, can itself be viewed as a family of coefficients of relational agreement, provided the appropriate uniforming transformation is applied before computing ICC(2,1) (Fagot, 1991); that is, a different estimate of ICC(2,1) can be computed for each possible scale type. Shrout and Fleiss (1979) point out that ICC(3,1) tends to be higher than ICC(2,1) and cite estimates of ICC(2,1) = .29 and ICC(3,1) = .71 for the data of their Table 2. When both indices are computed on the untransformed data, then the difference is attributable to the fact that ICC(3,1) treats additivity of ratings as meaningless disagreement (through removal of the main effect of judges) and ICC(2,1) treats additivity as meaningful disagreement among judges (attenuating the estimate). But ICC(3,1) is a coefficient of relational agreement only for additive scales, and if the two ICC coefficients are to be compared as coefficients of relational agreement, then the uniforming transformation for additive scales (1b) should be applied before computing ICC(2,1). If this is done for the Table 2 data of Shrout and Fleiss (1979), then ICC(2,1) = .75, slightly higher than ICC(3,1), with *both* ICC estimates treating additivity as meaningless disagreement.

The difference between ICC(2,1) and ICC(3,1) estimates for additive scales depends on the number of targets (n). With both ICC estimates computed on the uniformed versions (1b), JMS = 0, and it can be shown that

TABLE 2.

Examples of Lower Limits on Coefficients

	I				II				III			
Target	X_1	X_2	X_3	Mean	X_1	X_2	X_3	Mean	X_1	X_2	X_3	Mean
1	1	5	9	5	8	2	5	5	1	2	3	2
2	5	6	4	5	4	4	7	5	2	3	1	2
3	3	4	8	5	3	9	3	5	3	1	2	2
Mean	3	5	7	5	5	5	5	5	2	2	2	2
Coefficient												
I			-.20				-.50				-.50	
P			-.37				-.49				-.50	
A			-.50				-.50				-.50	
L			-.40				-.45				-.50	
ICC(3,1)			-.50				-.50				-.50	
ICC(2,1)			-.33				-1.00				-1.00	

$$\text{ICC}(2,1) = \frac{(n)\text{ICC}(3,1)}{\text{ICC}(3,1) + (n-1)}.$$

For the Shrout and Fleiss Table 2 data, the above formula gives ICC(2,1) = (6)(.71)/(.71 + 5.00) = .75, in agreement with the direct calculation. In general, for additive scales, $|\text{ICC}(2,1)| \geq |\text{ICC}(3,1)|$, the difference is controlled by n, and ICC(2,1) approaches ICC(3,1) as n approaches infinity.

Further Issues on Generalized Coefficients

Sign of Coefficients

In general, the sign of the coefficient for $k > 2$ should be interpreted in the same manner as a bivariate coefficient: A negative sign implies disagreement among judges, and a positive sign implies agreement among judges. When a bivariate coefficient reaches its lower limit of -1, it has a simple interpretation—a perfect negative relationship with high values of one variable paired with low values of the other variable. Obviously such a simple interpretation is not possible for more than two variables, but the following result shows that a similar interpretation can be made for negative coefficients when $k > 2$.

This result is that if target means are equal, then a systematic directional relationship does exist for negative coefficients, namely that *each judge variable is a perfect linear function, with negative slope, of the mean ratings of all other (k − 1) judges*:

$$r(X_i, \bar{X}_{\text{rem}}) = -1 \qquad (\bar{X}_{.1} = \bar{X}_{.2} = \cdots = \bar{X}_{.n} \text{ and } i = 1, \ldots, k), \tag{13}$$

where r is the Pearson bivariate linear correlation coefficient, $\bar{X}_{\text{rem}}$ is the mean rating of all judges *excluding* X_i, and $\bar{X}_{.t}$ is the mean of the t-th target.

For example, for Case I, Table 2, the mean ratings for the second and third judges are 7, 5, and 6 for Targets 1, 2, and 3, respectively. Hence, $(X_1, \bar{X}_{\text{rem}}) = (1,7), (5,5), (3,6)$, and for these three data points, $r(X_1, \bar{X}_{\text{rem}}) = -1$. Similarly, $r(X_2, \bar{X}_{\text{rem}}) = -1$ and $r(X_3, \bar{X}_{\text{rem}}) = -1$. As Case I of Table 2 shows, only coefficient A reaches a lower limit when target means are equal, although coefficients P and L are substantially negative. Thus, the directional interpretation of (13) holds even when I, P, and L do not reach a lower limit. Of course, the negative linear relationship of (13) will be approximate to the degree that inequalities in target means occur.

What (13) does accomplish is to make clear a sense in which negative coefficients for I, P, A, and L may be interpreted as directional, and not simply representing unsystematic disagreement among judges.

Disagreement Among Judges

The issue of disagreement among judges is complex. If interaction is large, a pattern of both positive and negative pairwise coefficients will occur, in which case the overall coefficient is relatively meaningless, both for this theory as well as for intraclass correlation. The situation is similar to ANOVA when attempting to interpret results in the presence of a significant interaction, which generally leads to further analysis of interaction effects. This suggests similar analyses of patterns of disagreement among judges. For example, hypothetical data from Hays (1988; problem #16, p. 541) was used to compute $L = .15$ ($k = 8$, $n = 5$), with fifteen of the pairwise coefficients negative and thirteen positive. Further analysis showed that the judges could be classified into two clusters, I and II, with I = (X_1, X_2, X_3, X_8) and II = (X_4, X_5, X_6, X_7). The linear coefficient for each cluster separately was large and positive, $L_{\text{I}} = .83$ and $L_{\text{II}} = .71$, with fifteen of the sixteen pairwise coefficients of linearity between judges in *different* clusters (e.g., r_{14}) *negative*. Hence, the analysis uncovered two clusters of judges, with strong judge agreement *within* each cluster and strong disagreement *between* clusters.

Choice of Coefficient

To appreciate the consequences of using an "inappropriate" coefficient, consider the use of the coefficient of Linearity if the scale is ratio, additive, or absolute. By

applying transform (1d), the coefficient L is an inflated estimate of P because (1d) treats additivity as meaningless disagreement; L is an inflated estimate of A because (1d) treats proportionality as meaningless disagreement; and L is an inflated estimate of I because (1d) treats both additivity and proportionality as meaningless disagreement. For example, for the Table 2 data of Shrout and Fleiss (1979), $I = .25$, $A = .72$, and $P = .65$. But $L = .76$, illustrating the inflation of the coefficient by erroneously treating additivity and/or proportionality as meaningless disagreement.

The choice of a coefficient of agreement based solely on scale type assumes that the "arbitrary assignments" determined by scale type are free to vary among judges (e.g., that judges are free to select their own units on a ratio scale). But if the arbitrary assignments (anchors) of a rating scale are set by the *experimenter*, then for perfect agreement, the ratings of all judges must be *identical* for all targets, and the coefficient of Identity is the appropriate coefficient. For example, consider a psychophysics task in which subjects make magnitude estimates (a) with modulus fixed by the experimenter (same for all subjects) and (b) no modulus.

For case (a) the same unit is assigned by the experimenter to all subjects, the ratings must be identical for all subjects, there is no meaningless disagreement, and hence the coefficient of Identity is the appropriate choice. For case (b), however, since subjects select their own unit, proportionality of ratings is meaningless disagreement, and P is the appropriate coefficient.

Summary

This article has presented a theory for assessing relational agreement for multiple judges based on the premise that the agreement coefficient depends on scale type, defined by the class of admissible transformations, and denotes agreement with respect to empirically meaningful relations. The relationship of the theory to intraclass correlation was examined, and it was shown that ICC(3,1) = A and hence that ICC(3,1) is a member of the proposed family of agreement coefficients.

Directions for future research could include extension of the theory to other scale types considered by Stine (1989a, Table 1). Stine (1989a) points out that there are other possible definitions of meaningful relations (see Narens, 1985, chap. 2.14), and such alternative definitions could be studied in relation to the assessment of agreement. Further study of the sampling properties of coefficients of agreement is needed (Zegers & ten Berge, 1985). And, finally, further investigation of the issue of judge disagreement, particularly involving patterns of positive and negative pairwise coefficients, is of major importance (see Lahey, Downey, & Saal, 1983).

Appendix: Proof that ICC(3,1) = A

Intraclass correlation Case 3: Two-way ANOVA model with random factor of targets and fixed factor of judges (Shrout & Fleiss, 1979).

$$\text{ICC}(3,1) = \frac{\text{BMS} - \text{EMS}}{\text{BMS} + (k-1)\text{EMS}}, \tag{A1}$$

where k = number of judges, n = number of targets, BMS = target mean square, and EMS = judge by target interaction mean square.

First it is shown that

$$\text{BMS} = \frac{\left[\sum_{i}^{k}\sum_{t}^{n} x_{it}^2 + 2\sum_{i<j}^{k}\sum_{t}^{n} x_{it}x_{jt}\right]}{k(n-1)}, \tag{A2}$$

where $x_{it} = X_{it} - \bar{X}_i$ for raw scores X. Sum of squares targets is $SS_B = k\,\Sigma_t\,(\bar{X}_{.t} - \bar{X})^2$, and since $\bar{X}_{.t} - \bar{X} = (1/k)\,\Sigma_i\,x_{it}$, then $SS_B = (1/k)\,\Sigma_t\,(\Sigma_i\,x_{it})^2$. It can be shown that $(\Sigma_i\,x_{it})^2 = \Sigma_i\,x_{it}^2 + 2\,\Sigma_{i<j}\,x_{it}x_{jt}$ and $\Sigma_t\,(\Sigma_i\,x_{it})^2 = \Sigma_i\,\Sigma_t\,x_{it}^2 + 2\,\Sigma_{i<j}\,\Sigma_t\,x_{it}x_{jt}$. Hence, $SS_B = (1/k)[\Sigma_i\,\Sigma_t\,x_{it}^2 + 2\,\Sigma_{i<j}\,\Sigma_t\,x_{it}x_{jt}.]$, and dividing by degrees of freedom $(n-1)$, BMS $= SS_B/(n-1)$ is given by A2.

Next it is shown that

$$\text{EMS} = \frac{\sum_{i}^{k}\sum_{t}^{n} x_{it}^2 - \left[\frac{2}{k-1}\right]\sum_{i<j}^{k}\sum_{t}^{n} x_{it}x_{jt}}{k(n-1)}. \tag{A3}$$

Since interaction constants $(\hat{\gamma}_{it})$ are equal for raw scores and deviation scores, one has $\hat{\gamma}_{it} = x_{it} - \bar{x}_{.t} - \bar{x}_{i.} + \bar{x}$, and since $\bar{x}_{i.} = 0$ and $\bar{x} = 0$, then $\hat{\gamma}_{it} = x_{it} - \bar{x}_{.t}$, and $\Sigma_i\,\Sigma_t\,\hat{\gamma}_{it}^2 = \Sigma_i\,\Sigma_t\,x_{it}^2 + k\,\Sigma_t\,\bar{x}_{.t}^2 - 2\,\Sigma_i\,\Sigma_t\,\bar{x}_{.t}x_{it}$, which simplifies to $\Sigma_i\,\Sigma_t\,\hat{\gamma}_{it}^2 = [(k-1)/k]\,\Sigma_i\,\Sigma_t\,x_{it}^2 - (2/k)\,\Sigma_{i<j}\,\Sigma_t\,x_{it}x_{jt}$, and dividing $\Sigma_i\Sigma_t\,\hat{\gamma}_{it}^2$ by interaction degrees of freedom $(n-1)(k-1)$, one gets (A3). Finally, substituting (A2) and (A3) into A(1) one gets ICC(3,1) $= A$.

References

Adams, E. W., Fagot, R. F., & Robinson, R. E. (1965). A theory of appropriate statistics. *Psychometrika, 30*, 99–127.

Alper, T. M. (1985). A note on real measurement structures of scale type $(m, m+1)$. *Journal of Mathematical Psychology, 29*, 73–81.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401–485). Hillsdale, NJ: Erlbaum.

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

Fagot, R. F. (1991). Reliability of ratings for multiple judges: Intraclass correlation and metric scales. *Applied Psychological Measurement, 15*, 1–11.

Fagot, R. F., & Mazo, R. M. (1989). Association coefficients of identity and proportionality for metric scales. *Psychometrika, 54*, 93–104.

Fisher, R. A. (1924). The influence of rainfall on the yield of wheat at Rothamsted. *Philosophical Transactions of the Royal Society of London, Series B, 213*, 89–124.

Green, B. F. (1956). A method of scalogram analysis using summary statistics. *Psychometrika, 21*, 79–88.

Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics (Vol. 2)*. London: Griffin.

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlation: There's more there than meets the eye. *Psychological Bulletin, 93*, 586–595.

Luce, R. D., Krantz, D. H., Suppes P., & Tversky, A. (1990). *Foundations of Measurement, Vol. 3: Representation, axiomatization, and invariance*. New York: Academic Press.

Narens, L. (1981a). A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory and Decision, 13*, 1–70.

Narens, L. (1981b). On the scales of measurement. *Journal of Mathematical Psychology, 24*, 249–275.

Narens, L. (1985). *Abstract measurement theory*. Cambridge, MA: MIT Press.

Roberts, F. S. (1979). *Measurement theory*. Reading, MA: MIT Press.

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

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (p. 1–49). New York: Wiley.

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science, 161*, 849–856.

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

Stine, W. W. (1989b). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin, 105*, 147–155.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, Vol. I* (pp. 1–76). New York: Wiley.

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*, 559–562.

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

Manuscript received 3/4/92
Final version received 7/7/92