Coefficients for Interrater Agreement

Frits E. Zegers
University of Groningen

The degree of agreement between two raters who rate a number of objects on a certain characteristic can be expressed by means of an association coefficient (e.g., the product-moment correlation). A large number of association coefficients have been proposed, many of which belong to the class of Euclidean coefficients (ECs). A discussion of desirable properties of ECs demonstrates how the *identity coefficient* and its generalizations, which constitute a family of ECs, can be used to assess interrater agreement. This family of ECs contains coefficients for both nominal and non-nominal (ordinal and metric) data. In particular, it is pointed out which information contained in the data is accounted for by the various coefficients and which information is ignored. *Index terms: association coefficients, correlation, Euclidean coefficients, generalized identity coefficients, interrater agreement.*

It often occurs that a number of persons or objects are rated on a certain characteristic by two or more judges (e.g., ratings of social helplessness of mentally disabled people, the quality of songs, or the quality of scientific papers). Consequently, there are two or more judgments or *scores* for each judged entity. Moreover, judges generally will not agree completely, unless the judgment task is trivial. Of course, a certain degree of discrepancy between judges cannot be avoided, but when there is a large amount of discrepancy, the value of the judgment procedure will be doubtful. Thus, the question arises of how to assess the degree of agreement between judges. This agreement is often expressed by means of the product-moment correlation (PMC), also known as Pearson's *r*. Two properties of the PMC are related to the idea of agreement: If the two judges produce two identical sets of scores, the PMC will attain its maximum value (+1), and if the judgments are randomly made, the PMC will be approximately zero.

In many situations, however, the PMC is not the proper measure of agreement. If the scores are *qualitative* (nominal data), for example, the PMC cannot be computed. In addition, some properties of the PMC may be undesirable in a given situation. Consider two teachers who grade the papers of three students on a 10-point scale ranging from 1 (very poor) to 10 (excellent), and scores of 6 or higher are considered "sufficient" and scores of 5 or lower are considered "insufficient." If one teacher gave grades of 7, 8, and 9, and the other teacher gave grades of 2, 3, and 4 to the same papers, the PMC between these two sets of three scores would be +1. The teachers did not fully agree, however. They agreed about the relative positions of the three papers, but they did not agree in an absolute sense about the quality of the papers.

The PMC is one of the many association coefficients that can be used in such an instance to assess agreement between judges. This paper explores the question of how to select an appropriate association coefficient, given a specific judgment task. After a discussion of the class of E coefficients (ECs), which is an important class of association coefficients, a family of association coefficients belonging to this class of ECs is discussed. Coefficients for non-nominal data (i.e., data with at least ordinal information) are described, and the family of ECs is generalized for nominal (categorical) data.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 4, December 1991, pp. 321-333

0146-6216/91/040321-13$1.90

E Coefficients

The association coefficients that will be discussed here belong to the class of ECs (Janson & Vegelius, 1978b)—the E referring to Euclidean. A specific association coefficient is an EC if the *association matrix* has certain characteristics, which are discussed below. For a given number of variables, the association matrix contains the association coefficients of all pairs of variables. A well-known example is the correlation matrix (with PMCs). In order to be an EC, the association coefficient itself does not need to be a PMC. In principle, however, variables should exist that have the association matrix as their correlation matrix with PMCs. These hypothetical variables need not have a simple relation to the original variables.

If an association coefficient is an EC, it will have a number of properties:

1. The association coefficient is symmetric (i.e., the association between variables X and Y equals the association between Y and X);
2. The maximum value of the association coefficient is 1;
3. The association coefficient of a variable with itself is 1;
4. The value of the association coefficient is never less than -1; and
5. If the association between variables X and Y is perfect (+1), the association between X and an arbitrary third variable W equals the association between Y and W.

Property 5 implies the transitivity of perfect association: If the association between X and Y and between Y and W are both perfect, then the association between X and W is perfect.

ECs have a number of desirable properties. First, some of the properties of ECs mentioned above fit with the concept of agreement. The symmetry of an EC (Property 1) mirrors the idea that there generally is no reason to distinguish the agreement between Judges A and B from the agreement between Judges B and A. Properties 2 and 3 imply that no person can agree more with a judge than the judge him/herself. Second, in order to facilitate the interpretation of the value of an agreement measure, this measure should be normed on a bounded interval. ECs are normed on the interval [-1, +1] (see Properties 2 and 4).

Third, from the definition of an EC, it follows that the association matrix of an EC can be used in components analysis (e.g., to look for structure in judgment data). Moreover, such a matrix may be converted into a matrix with distances between the variables in Euclidean space (Gower, 1966); this distance matrix can be used in metric multidimensional scaling.

Often it is difficult to determine whether or not a certain association coefficient is an EC. Sometimes it can be shown that a coefficient lacks one or more properties of an EC, which implies that the coefficient does not belong to the class of ECs. A practical way of proving that an association coefficient is an EC is to show that the coefficient belongs to a family of ECs. Association coefficients can be classified into (partially overlapping) families of coefficients. Coefficients within a family have a common structure: They can be written as variations on one basic formula. If it can be proven that such a basic formula yields an EC, then, by implication, all members of the family belong to the class of ECs.

A general family of ECs has been proposed by Zegers (1986b). This family comprises various other families of association coefficients (e.g., the family proposed by Daniels, 1944, which includes Kendall's *tau* and Spearman's *rho*; the family of Cohen, 1969, with coefficients for profile comparisons; the families of Janson and Vegelius, 1978a, 1978b, 1979, 1982a, 1982b, which include coefficients for variables of mixed measurement level; the family of Zegers and ten Berge, 1985; and the family of Zegers, 1986a, with coefficients for metric scales).

Coefficients for Non-Nominal Data

The scores of a judge may be numbers or labels. If these numbers or labels refer to various categories without any order relation among the categories, the data are called nominal. In all other cases, the data are non-nominal. The choice of a coefficient to assess agreement depends partly on whether the data are nominal or non-nominal; therefore, the coefficients for these two types of data are discussed separately.

As argued above, two judges will rarely produce identical sets of judgments. Yet it is useful to analyze the circumstances under which various association coefficients indicate perfect agreement. In general, association coefficients belonging to the class of ECs attain their maximum value of 1 if the sets of scores of two judges are identical. Conversely, an association coefficient of +1 does not imply that the two sets of scores are identical. For example, a PMC of +1 implies that the two sets of scores are identical within a positive linear (or affine) transformation (e.g., the example above of two teachers grading a number of papers). The *identity coefficient* introduced by Zegers and ten Berge (1985) as an association coefficient for absolute scales is +1 *if and only if* the two sets of scores are identical.

The identity coefficient is based on the sum of squared differences between corresponding scores of the two judges. If the two judges completely agree, this sum is 0, and the more the judges disagree, the larger this sum will be. By subtracting the sum of squared differences from +1, an index is obtained that is +1 in the case of identical judgments, and is smaller the more the judges differ. In order to obtain a coefficient that cannot have a value smaller than -1, the sum of squared differences is divided first by a factor d before it is subtracted from +1. This factor d is the sum of the two sums-of-squared scores, and the identity coefficient is obtained. If the scores of the two judges is given by X_i and Y_i, respectively, with i denoting the ith judged object, then the identity coefficient (e_{xy}) is given by

$$e_{xy} = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}X_i^2 + \sum_{i=1}^{n}Y_i^2} \quad , \tag{1}$$

where n is the number of judged objects. An equivalent but computationally simpler formula is

$$e_{xy} = \frac{2\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2 + \sum_{i=1}^{n}Y_i^2} \quad . \tag{2}$$

The identity coefficient belongs to the class of ECs (Zegers & ten Berge, 1985). The problem of assessing the agreement between the two teachers mentioned above now seems solved: Contrary to the PMC, the identity coefficient is not +1, but is smaller (.66), which indicates imperfect agreement. Suppose, however, that the scores only serve to give first, second, and third prizes to the three students. In this case, the usual meaning of the school grade scale is irrelevant; only the rank order of the scores given by the judges is of importance. The nonperfect identity coefficient now wrongly suggests disagreement among the teachers with respect to the distribution of the prizes. Clearly, a rank correlation coefficient in this situation is a better method for assessing agreement.

Stine (1989) argues that only certain relationships among the scores of one judge are empirically meaningful, and that other relationships are irrelevant. An agreement coefficient should express the agreement between the judges in terms of the empirically meaningful relationships among the scores, and it should ignore the (dis)agreement with respect to the irrelevant relationships. The empirically meaningful relationships are termed here *meaningful information*, and the irrelevant relationships are termed *irrelevant information*. According to Stine (1989), the scale type (measurement level) of the data determines which relationships among the data are empirically meaningful and which are irrelevant. It is impossible, however, to prove in a formal way the scale type of judgment data. It will be impossible, therefore, to determine on this basis which information will be meaningful and which will be irrelevant. A pragmatic approach to the issue of meaningful information is proposed below.

Judgment data are collected with a certain purpose. This purpose determines which information is meaningful and which is irrelevant. In one of the examples above, school grades were collected in order to distribute prizes to the students; thus, only the ordinal information was meaningful. If the school grades were collected in order to make pass/fail decisions, the meaningful information would be whether a score was above or below the pass/fail criterion. The question of how reliably the judgment procedure yields meaningful information can be answered by computing an appropriate agreement coefficient (i.e., a coefficient that takes into account only the meaningful information, and ignores the irrelevant information). In Stine's approach, the meaningful information results from the measurement level; in the pragmatic approach advocated here, the measurement level is implied by the identification of the meaningful information.

The concept of meaningful information can be used in situations in which an agreement coefficient should attain its maximum value of +1: An agreement coefficient should attain the value +1 if and only if the sets of scores of the two judges are identical in terms of the meaningful information. One method of finding a coefficient that satisfies this demand is to transform the scores of each judge to a *meaningful version* and then compute the identity coefficient between these meaningful versions. With this method, the problem of selecting an appropriate agreement coefficient is replaced by the problem of determining the meaningful versions of the scores given by the judges.

Transforming scores to their meaningful version consists of removing irrelevant information while preserving meaningful information, in order to obtain some kind of standardized version of the scores. A simple example of such a transformation is the standardization of variables, which results in standard scores with mean 0 and variance 1. Standardization preserves the interval information, but it removes information about the mean and the scaling. Zegers and ten Berge (1985) provide another example of such a transformation, including standardization as a special case (i.e., the *uniforming* of variables).

The construction of the meaningful version is performed here in three steps: (1) rank ordering or not rank ordering, (2) subtracting or not subtracting a reference point, and (3) rescaling or not rescaling. These steps are described in detail below; steps 2 and 3 use the scores resulting from the previous step(s).

1. If the only meaningful information is rank order information, the scores are replaced by their rank orders.
2. The scores are expressed as differences to a reference point by subtracting the value of the reference point from the scores. The reference point may be *absolute* or *relative.* An absolute reference point is sample-independent, which means that it does not depend on the observed scores. An absolute reference point may be thought of as a type of natural zero point, or a neutral point of the judgment scale (e.g., pass/fail point on a grade scale). A relative reference point depends on the observed scores (e.g., the mean or another sample-dependent measure of central tendency). The sample mean is the only relative reference point that is considered here.
3. If the *scaling* of the scores is arbitrary, the scores are rescaled to obtain a mean squared score equal to 1. This is done by dividing the scores by the square root of the mean squared score. Equation 2 shows that the value of the identity coefficient is not affected by multiplying or dividing both X and Y scores by the same constant. Therefore, rescaling as described here affects the resulting identity coefficient only if X and Y scores have different mean squares after step 2. Such a difference in mean squares may be the result of different rating styles of the two judges: one judge may be inclined to use more extreme points of the rating scale than the other. Whether a difference in scaling is considered meaningful depends on the specific judgment task. If it is not, the scores of both judges should be rescaled as described here

to remove the irrelevant scaling difference.

The three steps described above yield the meaningful versions of the scores. Inserting the meaningful versions in the formula of the identity coefficient (Equations 1 or 2) yields the desired agreement coefficient, which is in principle a specific agreement coefficient for each combination of the three steps. These coefficients are presented in Table 1; these coefficients were derived by examining the conditions in which agreement coefficients should be maximal (+1).

The combination of absolute reference point with rescaling and with reference point c, yields Cohen's (1969) r_c coefficient. In the special case of $c = 0$, this coefficient is identical to Tucker's (1951) congruence coefficient. The combination of absolute reference point with no rescaling and with $c = 0$ yields the identity coefficient (e) of Zegers and ten Berge (1985). By analogy with the r_c coefficient, the coefficient that results from the combination of absolute reference point with no rescaling and with $c \neq 0$ is denoted by e_c or c identity.

The combination of relative reference point with rescaling yields ordinary standard scores (with mean 0 and variance 1) as meaningful versions, with the PMC as the resulting coefficient. The combination of relative reference point with no rescaling yields the additivity coefficient (Zegers & ten Berge, 1985).

The scores will not be rescaled if a difference in original scaling should result in an agreement coefficient with a value less than +1. In the case of rank orders, a scaling difference indicates that the two sets of rank orders have different numbers of ties, or ties with different numbers of elements. In these cases, both rescaling and not rescaling will yield a coefficient with a value less than +1. Rescaling yields well-known coefficients—namely Spearman's rank correlation (ρ) and r_{oz} of Vegelius (1976; see Zegers, 1986b, pp. 42, 43)—and not rescaling yields two new coefficients with no apparent advantages.

Negative Agreement and Zero Agreement

All coefficients in Table 1 change sign without changing their absolute value if the scores of one judge are reflected with respect to the reference point. It can be verified easily that such a reflection changes the sign of the meaningful scores, which results in a sign change of the coefficient (see the numerator in Equation 2). With a relative reference point (the mean), a sign change of the scores is identical to reflection with respect to the reference point, because a sign change of the scores also changes the sign of the relative reference point. With an absolute reference point of 0 ($c = 0$), changing the sign of the scores is clearly identical to reflection. The sign change of an agreement coefficient as described here yields an interpretation of negative agreement or an *anti-agreement.*

In the case of coefficients based on scores with a relative reference point (the PMC and addi-

Table 1
Agreement Coefficients for Non-Nominal Data

	Reference Point		
	Relative (Mean)	Absolute c	
		$c = 0$	$c \neq 0$
No Rank Ordering			
Rescaling	PMC	Congruence (Tucker, 1951)	Cohen's r_c (Cohen, 1969)
No Rescaling	Additivity (Zegers & ten Berge, 1985)	Identity (Zegers & ten Berge, 1985)	c Identity
With Rank Ordering (for absolute c, c is the rank of the neutral point)			
Rescaling	Spearman's Rho	r_{oz} (Vegelius, 1976)	
No Rescaling	Not Recommended	Not Recommended	

tivity coefficient), a second interpretation of negative agreement results. The PMC (r_{xy}) can be expressed as

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad , \tag{3}$$

and the additivity coefficient (a_{xy}) can be expressed as

$$a_{xy} = \frac{2s_{xy}}{s_x^2 + s_y^2} \quad , \tag{4}$$

where s_{xy} denotes the covariance between the (original) scores, s_x and s_y denote the standard deviations, and s_x^2 and s_y^2 denote the variances of the (original) scores. It is well known that the covariance is 0 if the two variables are statistically independent. In the case of agreement scores, the covariance will have zero expectation if the judges produce scores in an unsystematic or random manner; this results in zero expectation of the PMC and the additivity coefficient. Negative agreement of these coefficients can be interpreted as less agreement than expected under chance.

For the coefficients in Table 1 that are based on scores with an absolute reference point, this second interpretation of negative agreement is not valid, because these coefficients cannot be expressed by means of a formula with only a covariance term in the numerator. Consider, for example, the identity coefficient for a rating scale with only positive scale points. It is clear from Equation 2 that the identity coefficient for this scale is guaranteed positive even if the judges produce scores randomly, within the restrictions of the rating scale. This demonstrates that a positive value of an agreement coefficient does not necessarily imply that the agreement is more than expected under chance. Therefore, it is important to determine the value of the coefficient under chance for a valid interpretation of the value of an agreement coefficient.

Agreement Coefficients Under Chance

Let the scores of two judges who judged n objects be given by X_i and Y_i, $i = 1, 2, \ldots, n$, which results in n pairs of scores (X_i, Y_i). If the pairing of the scores has been achieved randomly, then any other pairing of the X and Y scores will be as equally probable as the observed pairing. Thus for a fixed order of the X scores, each permutation of the Y scores is equally probable. There are $n!$ permutations, each resulting in a different pairing of the X and Y scores. For each of the $n!$ pairings, the value of the agreement coefficient can be computed. The value of the agreement coefficient under chance is defined as the mean of these $n!$ values. For the identity coefficient under chance ($\hat{e}_{xy}$) it is not necessary to actually compute these $n!$ values, as a simple formula can be derived:

$$\hat{e}_{xy} = \frac{2n^{-1}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} Y_i^2} \quad , \tag{5}$$

(see Zegers, 1986a). The interpretation of the observed value of the identity coefficient in Equation 2 can be compared with the value under chance in Equation 5.

The various coefficients in Table 1 were derived by using the formula of the identity coefficient with the meaningful versions of the scores. In like manner, the coefficient under chance can be determined by using the meaningful versions of the scores in Equation 5. It can be verified readily that the coefficients based on scores with a relative reference point equal to the mean have zero expectation under chance; under these circumstances, the numerator of Equation 5 contains the sums of deviation scores, which is 0. The comparison of the value of a coefficient (generally denoted as g_{xy}) and the value under chance ($\hat{g}_{xy}$) is based on the difference ($g_{xy} - \hat{g}_{xy}$). This difference can be used to construct a chance-corrected version of the coefficient.

Chance-Corrected Agreement Coefficients

A well-known method of chance correction of an association coefficient (g_{xy}) is to relate the difference of the coefficient and the coefficient under chance ($\hat{g}_{xy}$) to the theoretical maximum of this difference, which is ($1 - \hat{g}_{xy}$) for ECs. This yields a chance-corrected coefficient (g'_{xy}), given by

$$g'_{xy} = \frac{g_{xy} - \hat{g}_{xy}}{1 - \hat{g}_{xy}} \quad , \tag{6}$$

(see Zegers, 1986a, 1986b). If g_{xy} is an EC, then chance-correction by Equations 5 and 6 yields a chance-corrected coefficient g'_{xy}, which also belongs to the class of ECs. The result (i.e., a chance-corrected EC is an EC itself) is valid for the proposed method of chance correction, but it is not necessarily valid for other correction methods. Chance correction by Equation 6 does not affect the coefficients based on scores with a relative reference point (the mean), because the value under chance ($\hat{g}_{xy}$) is 0 for these coefficients. Whether or not corrected or uncorrected coefficients should be used in the case of scores with an absolute reference point depends on whether the chance correction as described here is adequate in the given situation.

It is important to note that for scores with an absolute reference point, the interpretation of negative agreement differs for uncorrected and corrected coefficients. Negative agreement for uncorrected coefficients can be interpreted as anti-agreement, because the sign of the coefficient changes if the scores of one of the two judges are reflected with respect to the reference point. The corrected coefficients do not have this property, because negative agreement for them has to be interpreted as less agreement than expected under chance. For coefficients based on scores with the mean as the relative reference point, both interpretations are valid, as is argued above; this agrees with the fact that these coefficients coincide with their chance-corrected versions.

Consider, for example, a situation with a large difference between corrected and uncorrected coefficients in which two teachers (X and Y) judge the quality of papers using a 10-point school grade scale. The scores are given in Table 2.

Table 2
Scores of Two Teachers (X and Y) for Four Papers Before and After Subtraction of Reference Point (5.5)

	Before		After	
Paper	X	Y	X	Y
1	8	8	2.5	2.5
2	8	9	2.5	3.5
3	9	8	3.5	2.5
4	9	9	3.5	3.5

The question of how well the two teachers agree can be answered in two ways. In one respect, the teachers show a high measure of agreement: Both judge the quality of the papers as "good" (8) or "very good" (9). This agreement is expressed by the identity coefficient, $e_{xy} = .997$. After subtracting the absolute reference point 5.5 from each score, as shown in Table 2, the value of the identity coefficient is still very high: $e_{xy} = .973$. Conversely, the teachers also show no agreement: A relatively low score of teacher X (8) is paired with both a relatively low score (8) and a relatively high score of teacher Y (9); and a relatively high score of teacher X goes with both a relatively low score and a relatively high score of teacher Y. This lack of agreement is properly expressed by the chance-corrected identity coefficient, $e'_{xy} = 0$. A closer investigation of the chance-correction method of Equations 5 and 6 may shed some light on these conflicting interpretations.

Although it was not stated explicitly, a null model was used above to derive a coefficient under chance; this model states that the observed scores show no agreement, except for agreement obtained by random factors. In fact, the null model underlying Equation 5 is relative or sample dependent: Given the observed scores or the observed marginal distributions, each specific pairing of X and Y scores is equally probable. If the agreement between the teachers in Table 2 is considered high (see the discussion above), then an absolute or sample independent null model should be used. Such an absolute null model can have various forms, depending on the supposed population distributions. In the case of an absolute null model, the scores of the judges are assumed to constitute a random sample from a specified distribution. If the samples of both judges are drawn independently, the coefficient under chance is the expectation of the coefficient.

Suppose, for example, that teachers grading papers with the 10-point scale generally produce scores with the following frequency distribution: 10% for Grade 4, 15% for Grade 5, 25% for Grade 6, 25% for Grade 7, 15% for Grade 8, and 10% for Grade 9. Using computer simulation, the expected value of the identity coefficient for four papers is .953, and the expected value of the identity coefficient using the absolute reference point 5.5 is .278. The identity coefficients obtained for the data of Table 2 can be compared with these expected values. Using Equation 6 yields chance-corrected coefficients with values .936 and .963 for the "before" and "after" ratings, respectively.

The values of the expected identity coefficients given above were computed with 200,000 samples of size 4. An alternative is to use results obtained by Fagot and Mazo (1989). Without making any distributional assumption, they derived an asymptotically correct formula for the expectation of the identity coefficient between two independent variables. This expectation is expressed in terms of population means and variances of both variables; these parameter values are available in the case of an absolute null model.

Even in the small example used here and the assumed distribution, the computer simulated values were close to the values computed by means of the Fagot and Mazo formula (.954 and .328, respectively). This result suggests that even with moderate sample sizes, the Fagot and Mazo formula may be used safely for computing the expected value of the identity coefficient in the case of an absolute null model. Using expected values based on an absolute null model in the formula of chance correction (Equation 6) has one serious drawback, however: The resulting corrected coefficient is not guaranteed to be an EC.

Assessing agreement between two judges with non-nominal scores, which is discussed above, can be summarized as follows: First, the scores of the judges are transformed to their meaningful versions, which may involve rank ordering, subtraction of an absolute or relative reference point, and rescaling. Next, the identity coefficient between these meaningful versions is computed. The obtained value may be compared with the expected value, using a relative (sample dependent) or absolute (sample independent) null model. A potential drawback of the identity coefficient is discussed below.

Gower's Coefficient

Suppose two judges X and Y rate four objects on a 5-point scale, ranging from 1 to 5, with an absolute reference point of 3 (the scale center). Two different sets of meaningful scores are given in Table 3, with values of 2/3 and 1/2 for the identity coefficient. After chance correction by Equations 5 and 6, these values are 49/81 and 1/2. The agreement apparently differs for the two sets of data. It can be argued, however, that the agreement between the two sets of data in this table is equal, because the differences between the judges for each single object are the same. The numerator of the identity coefficient in Equation 1 contains the sum of the squared differences, and these sums are equal for the two datasets. But the denominators differ because it is the total sum of squares. Therefore, the values of the identity coefficient are not equal.

Table 3
Meaningful Scores of
Two Sets of Judges

	Set 1		Set 2	
Object	X	Y	X	Y
1	2	1	2	1
2	1	2	0	1
3	0	1	-1	0
4	0	1	-1	0

A coefficient that does have equal values for the data in Table 3 has been developed by Gower (1971):

$$G_{xy} = 1 - \frac{\sum_{i=1}^{n} |X_i - Y_i|}{nR} \quad , \qquad (7)$$

where X_i and Y_i are scores of the two judges; n is the number of objects; and R is the range of the rating scale, that is, the maximum value of the absolute difference $|X_i - Y_i|$($R = 4$ in Table 3). For both sets of data in Table 3, $G_{xy} = .75$.

Gower (1971) showed that G_{xy} is an EC. It is obvious from Equation 7 that G_{xy} uses the sum of absolute differences, not the sum of squared differences. In order to obtain a normed coefficient, this sum of absolute differences is divided by the theoretical maximum of that sum, nR; this maximum is obtained if the judges differ maximally for each object. Under these circumstances, the judges use the opposite extremes of the rating scale for each object, which results in $G_{xy} = 0$. Obviously, G_{xy} can have values on the interval [0,1]. The main difference between Gower's coefficient and the identity coefficient is the way in which the coëfficients are normed. The identity coefficient is normed in a relative or sample-dependent manner by means of the sums of squares of the meaningful versions. Gower's coefficient, however, is normed in an absolute or sample-independent manner using the range of the rating scale.

As a result of the absolute method of norming, Gower's coefficient can be interpreted as a measure of average agreement between judges per object. If

$$G_{xy_i} = 1 - \frac{|X_i - Y_i|}{R} \qquad (8)$$

denotes the agreement between X and Y in terms of object i, then the mean of this object agreement (averaged over the n objects) is equal to Gower's coefficient.

Agreement Coefficients for Nominal Data

Nominal data are obtained if a judge has to place a number (n) of persons or objects into a number (k) of mutually exclusive categories without ordinal relationships between the categories. If two judges place the same n objects into categories, they can use the same set of categories or different sets. This results in two classes of agreement coefficients for nominal data.

The Same Categories

If two judges place the same n objects into the same k categories, the data can be represented by means of a bivariate frequency table, with k rows and k columns. An arbitrary entry (g, h) of this table contains the number of objects placed into category g by judge X and into category h by judge Y. An arbitrary diagonal entry (g, g) contains the number of objects placed into category g by both judges. The categories contain no ordinal information; therefore, the order in which the categories are displayed is arbitrary, but it must be the same for both judges in order to ensure that category g of judge X is the same as category g of judge Y. Table 4 represents the judgments of two judges X and Y who judged 10 objects ($n = 10$) using three categories A, B, and C ($k = 3$).

An obvious agreement coefficient is the proportion agreement (p_o), defined as the sum of the diagonal entries of the bivariate frequency table (shown for these data in Table 5) divided by n. For the data in Tables 4 and 5, $p_o = 5/10 = .50$. Obviously, the proportion agreement is bounded on the interval [0,1].

Table 4
Nominal Data Categorizations of 10 Objects by Judges X and Y

Object	X	Y
1	A	B
2	A	A
3	B	B
4	C	B
5	A	B
6	C	C
7	C	C
8	B	B
9	C	A
10	B	C

Table 5
Bivariate Frequency Table for the Data in Table 4

Judge	Judge Y			
X	A	B	C	Sum
A	1	2	0	3
B	**0**	2	1	3
C	1	1	2	4
Sum	2	5	3	10

The proportion agreement can be compared with the value under chance, which is determined as follows: A column with row sums is added to the bivariate frequency table. The *g*th element of this column is the number of objects placed into category *g* by judge X. This column will be called the marginal distribution of judge X. In like manner, the marginal distribution of judge Y can be added (as a row) to this table. Given these marginal distributions, the value under chance (expected frequency) can be computed for each entry (g, h) in the table. This expected frequency is the product of the *g*th element of the marginal distribution of *X* and the *h*th element in the marginal distribution of *Y*, divided by *n* (cf. the computation of χ^2). The proportion agreement under chance (p_e) is the sum of the expected frequencies on the diagonal, divided by *n*. For the data of Table 5, $p_e = (.6 + 1.5 + 1.2)/10 = .33$. Obviously, this computation is based on a relative or sample-dependent null model, given the observed marginal distributions.

Correcting the proportion agreement for chance yields Cohen's (1960) kappa coefficient, which is expressed as

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad , \tag{9}$$

with p_e as described above. For the data of Table 4, $\kappa = (.50 - .33)/(1 - .33) = .17/.67 = .25$. Both κ and various other agreement coefficients for nominal data with the same categories can be derived in a manner analogous to that used for the coefficients for non-nominal data described above (Gower's coefficient excluded). A short outline of this procedure follows, which is based on indicator matrices. For a more detailed description, see Zegers (1986b, pp. 45–53).

Nominal data of one judge can be represented by means of an indicator matrix with *n* (the number of objects) rows and *k* (the number of categories) columns. Entry (i, g) of this matrix is 1 if the judge placed object *i* into category *g*; otherwise, it is 0. Each row of the indicator matrix contains a single 1 and $k - 1$ 0s. The indicator matrices of the two judges of Table 4 are given in Table 6.

The indicator matrix may be interpreted as a quantification of the nominal data of one judge, without loss of information. This quantification enables the use of (adapted) coefficients for non-nominal data. The identity coefficient can be generalized to assess the amount of identity of two matrices of scores instead of the identity of two columns of scores. A necessary and sufficient condition is that corresponding rows and columns of the two matrices have the same meaning. The indicator matrices of two judges who judge the same *n* objects using the same *k* categories satisfy this condition.

The generalized identity coefficient for two matrices is obtained as follows: (1) the numerator of the last term of Equation 1 is replaced by the sum of squared differences of corresponding entries of the two matrices; and (2) the denominator is replaced by the sum of the sum of squared entries of the first matrix and the sum of squared

Table 6
Indicator Matrices for the Data of Table 4

	Judge X			Judge Y		
Object	A	B	C	A	B	C
1	1	0	0	0	1	0
2	1	0	0	1	0	0
3	0	1	0	0	1	0
4	0	0	1	0	1	0
5	1	0	0	0	1	0
6	0	0	1	0	0	1
7	0	0	1	0	0	1
8	0	1	0	0	1	0
9	0	0	1	1	0	0
10	0	1	0	0	0	1

entries of the second matrix (the same result is obtained by placing the columns of each matrix under each other and computing the ordinary identity coefficient between these long columns). This formulation demonstrates that the generalized identity coefficient, being the ordinary identity coefficient between long columns, is an EC. It can be verified that the generalized identity coefficient used with indicator matrices yields the proportion agreement (p_o). If X_i and Y_i (with $i = 1, 2, \ldots, 30$) denote the scores in the long columns derived from the indicator matrices in Table 6, then $\Sigma X_i Y_i = 5$ and $\Sigma X_i^2 = \Sigma Y_i^2 = 10$. The resulting value of the identify coefficient is $2 \times 5/(10 + 10) = 1/2 = p_o$ (the proportion agreement).

The method of deriving a coefficient under chance, based on permutations of the scores of one judge, also may be generalized by permuting rows of one of the two matrices. A permutation of rows is a specific ordering of the rows. With n rows in the indicator matrix, there are $n!$ different orderings or permutations of these rows. In the case of indicator matrices, this yields the proportion agreement under chance (p_e). This proves that κ in Equation 9 can be considered a chance-corrected generalized identity coefficient, which implies that κ is an EC (see also Zegers, 1986b, p. 47).

Non-nominal scores can be transformed into meaningful versions (as has been discussed above). Using the identity coefficient with these meaningful versions yields various specific coefficients. In an analogous way, indicator matrices may be transformed (e.g., by centering rows and/or columns). Using the generalized identity coefficient with specifically transformed indicator matrices yields specific ECs. Well-known coefficients which can be obtained in this way are the C and S coefficients of Janson and Vegelius (1979) and the simple matching coefficient (Sokal & Sneath, 1963; see Zegers, 1986b, pp. 59–60).

The concept of meaningful versions is not very useful with nominal data. In fact, all meaningful information is contained in the indicator matrix and no meaningful information is removed or added by a transformation of the indicator matrix. In practice, that transformation will be selected which leads to a coefficient with which the researcher is familiar, or which fits in the tradition of a particular field of research. The main advantage of expressing an agreement coefficient as a (chance-corrected) generalized identity coefficient is that this implies that the coefficient in question is an EC.

Different Categories

Sometimes judges must classify a number of objects in a number of nonoverlapping classes, the meaning of which has to be determined independently by each judge. The number of classes may be specified in advance, or it may be left to the judge. The nominal categories are the names or labels of the distinct classes. The proportion agreement cannot be computed with this kind of data, because the categories of one judge do not correspond with the categories of the other (i.e., the gth category of judge X does not match the gth category of judge Y). For the same reason, it makes no sense to compute the identity between the two indicator matrices.

One possible solution is to base the agreement coefficient on the number of pairs of objects that are placed in the same category by both the first and the second judge. Dividing this number by the total number of pairs of objects yields the *dot product* (see Popping, 1983, p. 96). This dot product can also be expressed as a generalized identity coefficient—not between (transformations of) indicator matrices, but between two object matrices. The object matrix of a judge is a square matrix with n rows and columns. Entry (i,j) is 1 if the judge placed objects i and j in the same category, and is 0 otherwise. It can be shown that the generalized identity coefficient between two object matrices is identical to the dot product.

The generalized identity coefficient can also be computed between transformations of the object matrices (see Zegers, 1986b, pp. 50–53). Depending on the type of transformation the following coefficients result: The T^2 coefficient

(Tschuprow, 1939); the gamma coefficient (Hubert, 1977); the *J* coefficient (Janson & Vegelius, 1982b); and the *I* coefficient (Saporta, 1975), which is identical to the *T* coefficient (Janson & Vegelius, 1978b). A detailed discussion of the *J* and *T* coefficients and their relation with the *C* and *S* coefficients (for nominal data with the same categories) can be found in Zegers and ten Berge (1986). As in the case of nominal data with the same categories, the choice of a specific coefficient does not depend on considerations with respect to meaningful information, but depends rather on preferences or tradition.

Conclusions

A large number of agreement coefficients have been discussed. The choice of a coefficient in a specific context depends on a number of implicit or explicit assumptions with respect to the nature of the data and chance correction.

This overview, however, is not at all complete. In particular, association coefficients for variables of mixed measurement levels have not been discussed, because such coefficients do not play an important role in the context of assessing agreement between judges. However, ECs which can also be expressed as generalized identity coefficients have been developed for variables of mixed measurement levels, (see Janson and Vegelius, 1982a; Zegers, 1986b, pp. 55–56; and Zegers & ten Berge, 1986). With such coefficients, it would be possible to reflect the association between variables such as income, educational level, gender, weight, and blood group in a single association matrix.

References

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

Cohen, J. (1969). r_c: A profile similarity coefficient invariant over variable reflection. *Psychological Bulletin, 71,* 281–284.

Daniels, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika, 36,* 129–135.

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

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika, 53,* 315–328.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27,* 857–871.

Hubert, L. (1977). Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology, 30,* 98–103.

Janson, S., & Vegelius, J. (1978a). *Correlation coefficients for more than one scale type and symmetrization as a method of obtaining them* (Research Rep. No. 78-2). Uppsala, Sweden: University of Uppsala, Department of Statistics.

Janson, S., & Vegelius, J. (1978b). *On construction of E correlation coefficients* (Research Report 78-11). Uppsala, Sweden: University of Uppsala, Department of Statistics.

Janson, S., & Vegelius, J. (1979). On generalizations of the *G* index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14,* 255–269.

Janson, S., & Vegelius, J. (1982a). Correlation coefficients for more than one scale type. *Multivariate Behavioral Research, 17,* 271–284.

Janson, S., & Vegelius, J. (1982b). The *J* index as a measure of nominal scale response agreement. *Applied Psychological Measurement, 6,* 111–121.

Popping, R. (1983). *Overeenstemmingsmaten voor nominale data (Agreement coefficients for nominal data).* Doctoral dissertation. University of Groningen, The Netherlands.

Saporta, G. (1975). *Liaison entre plusieurs ensembles des variables et codage de donnés qualitatives (Relation between various sets of variables and the coding of qualitative data).* Doctoral dissertation. University of Paris, France.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy.* San Francisco: Freeman & Co.

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

Tschuprow, A. A. (1939). *Principles of the mathematical theory of correlation.* New York: William Hodge.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Report No. 984). Washington DC: Department of the Army, Personnel Research Section.

Vegelius, J. (1976). On generalizations of the *G* index. *Educational and Psychological Measurement, 36,* 595–600.

Zegers, F. E. (1986a). A family of chance-corrected association coefficients for metric scales. *Psychometrika, 51,* 559–569.

Zegers, F. E. (1986b). *A general family of association coefficients. Groningen, Netherlands: Boomker.

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

Zegers, F. E., & ten Berge, J. M. F. (1986). Correlation coefficients for more than one scale type: An alternative to the Janson and Vegelius approach. *Psychometrika, 51,* 549–557.

Acknowledgments

The author thanks J. M. F. ten Berge for his helpful comments on various versions of the manuscript and an anonymous reviewer for pointing out a result obtained by Fagot and Mazo (1989). This article is an adaptation of Het meten van overeenstemming, *which appeared in the* Nederlands Tijdschrift voor de Psychologie, 44, *(1989), 145–156. A computer program for most of the coefficients discussed will be published by ProGamma: Groningen, The Netherlands.*

Author's Address

Send requests for reprints or further information to Frits E. Zegers, University of Groningen, Department of Psychology, Grote Kruisstraat 2/I, 9712 TS Groningen, The Netherlands.