Can One Use Cohen's Kappa to Examine Disagreement?

Alexander von Eye[1] and Maxine von Eye[2]

[1]Michigan State University
[2]University of Bath

Abstract. This research discusses the use of Cohen's κ (kappa), Brennan and Prediger's κ_n, and the coefficient of raw agreement for the examination of disagreement. Three scenarios are considered. The first involves all disagreement cells in a rater × rater cross-tabulation. The second involves one of the triangles of disagreement cells. The third involves the cells that indicate disagreement by one (ordinal) scale unit. For each of these three scenarios, coefficients of disagreement in the form of κ equivalents are derived. The behavior of the coefficients of disagreement in the three situations is studied. The first and the third case pose no particular problem. The κ equivalents and the other coefficients can be interpreted as usual. In the second case, problems arise such that the range of disagreement κs is restricted because the tables are incomplete. Thus, the standard log-frequency model of rater independence is no longer applicable. When the more general models of quasi-independence are used, negative degrees of freedom can result for smaller tables. Simulation results illustrate the characteristics of the coefficients of disagreement for each of the three scenarios. Empirical data examples are given.

Cohen's (1960) κ (kappa) is the most popular coefficient of rater agreement. κ indicates the degree to which two raters agree beyond chance. It is a summary measure of agreement in a rater × rater cross-classification. Researchers often ask additional questions concerning agreement tables. Examples of such questions concern patterns of disagreement. These questions are of importance in rater training or when disagreement can have negative implications, for instance, for patients. In this article, we propose derivatives of Cohen's κ for the analysis of such questions. The benefit from such coefficients is that interpretation can follow the lines of κ, and no additional line of thinking or comparisons of results based on different statistical models are needed.

Analyzing Agreement Tables

Four approaches to analyzing cross-classifications of raters' judgements, that is, agreement tables, have been pursued (von Eye & Mun, 2005). The first approach involves estimating summary coefficients such as Cohen's coefficient κ (1960), raw agreement, Cohen's weighted κ, Brennan and Prediger's κ_n (1981), von Eye's (2005) κ alternative, or intraclass correlations (for continuous variables). Most of these coefficients share the characteristics that (1) they summarize the information in the table in one coefficient, (2) they are easy to interpret, typically as proportionate reduction-in-error (PRE) coefficients, and (3) significance tests exist that allow one to determine whether agreement occurred beyond what was expected based on a chance model.

The second approach to analyzing agreement tables involves estimating models. Two lines of research have been developed. The first line focuses on manifest variable models. von Eye and Mun (2005) proposed a family of log-linear models that not only allow one to focus on overall agreement but also on more detailed hypotheses that concern, for instance, specific weights that raters may have used (cf. Tanner & Young, 1985), scale characteristics of the rating scales as nominal or ordinal, the effect of covariates, or classifications of judgements (see also Agresti, 2002; Fleiss, Levin, & Paik, 2003; Schuster & Smith, 2002; Schuster & von Eye, 2001).

The second line of research involves the development of latent variable and mixture models (cf. Agresti, 1992; Schuster, 2002; Uebersax, 1993). For example, Agresti (1992) proposed latent class models for the agreement among three raters. This model assumes an unobserved categorical variable. The judgements in the categories of this variable are homogeneous, and the judgements of the three raters are statistically independent at each of the levels of the latent variable.

Latent trait models have been proposed, for example, by Wolfe (2004). These models allow one to test hypotheses concerning rater effects, for instance, rater characteristics, rater-specific accuracy and inaccuracy, and rater-specific trends such as centrality versus extremism. Extensions of latent variable models consider,

Methodology 2005; Vol. 1(4):129–142
DOI 10.1027/1614-1881.1.4.129

for instance, continuous latent variables (Uebersax, 1993) or Rasch models (cf. Lindsay, Clogg, & Grego, 1991).

These approaches are most flexible, and they allow one to test a large number of hypotheses about the structure of agreement tables. However, these approaches have rarely been used. Two reasons why they are rarely used may be that specific hypotheses do not exist very often and that model specification may not always be straightforward, from a user's perspective. For example, the number of parameters that need to be estimated for latent class models for rater agreement is typically large and models are often not identified. The number of parameters can be reduced by setting parameters to zero or by setting parameters equal. Placing such constraints is not always trivial, and results depend on these decisions. Users often do not entertain hypotheses that correspond to the required decisions. Another downside to these models is that they are not always applicable. For example, for binary rating variables, three or more raters are needed for a latent variable model to be identified. Accordingly, in log-linear models, degrees of freedom can be negative as well, for example in the stratified models proposed by Graham (1995). Here again, decisions need to be made and justified to reduce the number of parameters in a model.

The third approach involves exploring agreement tables. von Eye and Mun (2005) proposed using Configural Frequency Analysis (CFA; von Eye, 2002) to examine individual cells of agreement tables, and to ask whether the frequencies in these cells differ from what was expected under some probability model. CFA models allow one to search for agreement types (over-frequented main diagonal cells), agreement antitypes (under-frequented diagonal cells), disagreement types (over-frequented off-diagonal cells), and disagreement antitypes (under-frequented off-diagonal cells). Covariates can be considered as well as special effects. In addition, models for ordinal variables can be estimated. The CFA approach is straightforward and easily implemented. However, it does not lead to a summary statement about characteristics of the agreement table.

A fourth approach was recently proposed by DeCarlo (2002, 2005). The author presented a latent class extension of signal detection theory that allows one to test theories of psychological processes that underlie raters' behavior. Measures of the precision of raters and the accuracy of classifications can be calculated. To the best of our knowledge, there have been no applications of this approach yet, except DeCarlo's own (2005).

For the reasons listed, applications of modeling approaches to the analysis of rater agreement have been rare, although models have been discussed for more than 25 years. In addition, users occasionally question the benefit from such models. For example, Tanner and Young's (1985) model contains a parameter that expresses strength of agreement, and Schuster and Smith's (2002) contains a parameter that can be interpreted as the proportion of systematic agreement. Obviously, the context of a model can make this kind of parameter more informative than plain coefficients. This benefit, however, seems to escape a good number of users, and the coefficients of rater agreement remain discussed in the methodological literature, and employed in the applied literature far more often than models.

In the present article, we look again at Cohen's κ. Specifically, we ask whether κ can be used to create summary statements about selections of cells in the table that indicate disagreement. If this is possible, specific hypotheses in addition to those concerning PRE-type agreement can be formulated and tested. In the following section, we review Cohen's κ, the coefficient of raw agreement *ra,* and Brennan and Prediger's (1981) κ_n. Then, we present hypotheses concerning the off-diagonal cells of agreement tables, we propose corresponding κ derivatives, and present simulation results that indicate whether these coefficients can be used to test hypotheses other than those concerning perfect agreement. These simulations concern both the coefficients and statistics used for significance testing.

Coefficients of Agreement

In this article, we consider standard agreement tables for two raters. These are square, $I \times I$ tables, the I rows of which represent the rating categories used by Rater A, and the I columns of which represent the rating categories used by Rater B. These categories are typically the same. The entries in an agreement table, m_{ij}, indicate the observed frequencies of agreement–disagreement patterns. The probabilities of these patterns are p_{ij}.

Cohen's κ

In the case of two raters, Cohen's κ relates the probabilities in the agreement cells, these are the cells with index *ii,* to the expected cell probabilities in these cells. Specifically, let the sum of the probabilities in the agreement cells be

$$\theta_1 = \sum_{i=1}^{I} p_{ii}$$

where I is the number of rating categories used by both raters. Under the assumption of rater independence, the corresponding sum of expected probabilities in agreement cells is

$$\theta_2 = \sum_{i=j} p_{i.}\, p_{.j}$$

with $i, j = 1, \ldots, I$, and $i.$ indicating the ith row total, and $.j$ indicating the jth column total. The relation between θ_1 and θ_2 describes the degree to which agreement deviates from chance. If $\theta_1 > \theta_2$, the two raters agree to a degree that exceeds what was expected based on chance.

To express Cohen's κ in a standardized form, we relate the difference between θ_1 and θ_2 to the maximum difference between θ_1 and θ_2, that is, to $1 - \theta_2$. The difference $\theta_1 - \theta_2$, weighted by this maximum difference, is known as Cohen's (1960) κ,

$$\kappa = \frac{\theta_1 - \theta_2}{1 - \theta_2}.$$

κ describes the portion of judgements in which the two raters agree exactly, relative to the expected portion. κ can be estimated for a sample by using the observed cell frequencies.

Fleiss, Cohen, and Everitt (1969; cf. Hildebrand et al., 1977) showed that, for sufficiently large samples, the estimate of κ is normally distributed with mean κ. The authors derived the standard error of the estimate of κ. Using this standard error, the test statistic $z_\kappa = \hat{\kappa}/\hat{\sigma}_\kappa$ can be used to test the null hypothesis that $\kappa = 0$. As an alternative to this z test, von Eye and Brandtstädter (1988) proposed using the binomial test.

Characteristics of Cohen's κ.

We now review three of the well known characteristics of Cohen's κ, including the occasionally critical discussion of the coefficient (cf. von Eye & Mun, 2005).

First, the range of κ is $-\theta_2/(1 - \theta_2) \leq \kappa \leq 1$. $\kappa = 0$ if the probability of disagreement is the same as the probability of agreement, κ can be zero even if raters' judgements are not independent, and $\kappa = 1$ only if the probability of disagreement is zero.

Second, if the probability of disagreement is nonzero, the maximum value of κ decreases as the marginals deviate from a uniform distribution. This characteristic has been termed marginal dependency or prevalence dependency of chance-corrected agreement (Guggenmoos-Holzmann, 1995). It is interesting to note that large parts of the literature view this characteristic as a negative, and take it as a reason to recommend other methods of assessment of rater agreement. However, there are two trends in the recent literature that point to the positive aspects of this characteristic. The first trend is that authors suggest viewing the appraisal of the degree of agreement as base model-dependent. Examples of this approach include the work of von Eye and Sörensen (1991; cf. von Eye & Mun, 2005) who contrast chance models. The second trend is the repeated attempt to devise models that include parameters that have the same interpretation as κ. Examples of this work include the models of Tanner and Young (1985) and Schuster (2001).

Third, negative values of κ, quantifying disagreement, have the same interpretation as positive values. However, κ does not show monotonic behavior when the probability of disagreement is higher than the probability of agreement (von Eye & Sörensen, 1991).

Fleiss (1981) proposes, as rules of thumb, that values of $\kappa < 0.4$ indicate poor agreement, values of $0.4 \leq \kappa \leq 0.75$ indicate good agreement, and larger values indicate excellent agreement (cf. Landis & Koch, 1977).

Cohen's Weighted κ

The original coefficient κ places equal weights on all disagreement patterns. Thus, it does not allow one to distinguish between minor discrepancies, by say, 1 scale point, and major discrepancies, by more than 1 scale point. To allow such a distinction, Cohen (1968) introduced *weighted* κ for ordinal scales, or for nominal scales in which different types of disagreement come with different implications. We define, in a fashion parallel to the previous one,

$$\theta_1^* = \sum_i \sum_j \omega_{ij}\, p_{ij}$$

and

$$\theta_2^* = \sum_i \sum_j \omega_{ij}\, p_{i.}\, p_{.j},$$

where the probabilities are defined as before, and the ω_{ij} are the weights. Cohen specified that these weights (1) range $0 \leq \omega \leq 1$, and (2) be ratios. The second specification implies that a score of $\omega = 1$ carries twice the weight of a score of $\omega = 0.5$. As for κ, one can define weighted κ by

$$\kappa_w = \frac{\theta_1^* - \theta_2^*}{1 - \theta_2^*}.$$

For the standard error of κ_w, see Fleiss, Cohen, and Everitt (1969).

Brennan and Prediger's κ_n

One of the major criticisms of Cohen's κ is that its maximum value is less than 1.0 if both (1) the marginal distributions are not uniform, and (2) not all off-diagonal probabilities are zero. To remedy this problem, Brennan and Prediger (1981) suggested replacing the main effect model of rater independence by the null model of a uniform response distribution. The resulting coefficient of rater agreement is then

$$\kappa_n = \frac{\theta_1 - 1/I}{1 - 1/I},$$

where I is the number of categories that both raters used. As for κ, κ_n can be estimated from the observed frequencies in an agreement table.

The coefficient κ_n is sensitive to deviations from a null model (uniform distribution) as well as to deviations from a main effect model. It indicates the magnitude of agreement instead of agreement beyond chance. κ_n can reach its maximum value of 1.0 even if the marginal probabilities are not uniform. The magnitude of the coefficient itself serves as an indicator of the type of deviation only if a table possesses marginal homogeneity. This coefficient has recently been criticized because it can be large when raters randomly assign cases to rating categories while using these categories at different base rates (Hsu & Field, 2003; for significance testing and distributional characteristics see von Eye & Mair, 2005).

The Coefficient of Raw Agreement

The coefficient of raw agreement indicates the probability for two or more raters' judgements to match perfectly. This corresponds to the probability for two judgements to be located in the main diagonal of an agreement matrix. More specifically, *ra,* the coefficient of raw agreement, is

$$ra = \sum_i p_{ii} = \theta_1.$$

This coefficient can also be estimated from the observed frequencies by substituting p_{ii} with m_{ii}/N, where m_{ii} is the frequency observed for cell *ii* and *N* is the total number of judgements in an agreement matrix.

Often, researchers report both κ_n and *ra.* However, there is no need to report both κ_n and *ra* because, as was shown by von Eye and Mair (2005), the two measures are linear transformations of each other such that

$$\kappa_n = -\frac{1}{I-1} + ra\left(1 + \frac{1}{I-1}\right).$$

Three Questions Concerning Disagreement Cells

In this section, we discuss three questions concerning disagreement cells, that is, the off-diagonal cells in an agreement table. For each of these questions, we ask whether it can be answered by applying κ, κ_n, and *ra* to the cells addressed by these questions.

To introduce these questions and to illustrate the use of the coefficients reviewed in the previous sections, we present a data example (adapted from Schuster & Smith, 2002; Table 1). In a study on the agreement between service facility and research diagnoses of 223 psychiatric patients, patients were classified into four mutually exclusive ordered diagnostic categories. We name these (1) severe psychosis, (2) average severity psychosis, (3) mild psychosis, and (4) no clinical diagnosis. Table 1 gives the cross-classification of the facility with the research diagnoses.

Raw agreement between the two diagnostic procedures is 58.74%. This is 43.15% better than chance (κ). The improvement over the chance model of independence of diagnoses is significant ($z_\kappa = 10.48; p < 0.01$). Brennan and Prediger's (1981) κ_n assumes the value of 0.45. From these results, we conclude that the two diagnostic procedures represent judgements that are significantly in better agreement than expected based on chance for both the independence and the null models.

We now discuss questions researchers may have in addition to rater agreement. Specifically, we discuss three questions concerning disagreement. The first question is whether the off-diagonal cells, both above and below the main diagonal, allow one to make statements about whether significant lack of disagreement exists. The second question concerns the disagreement cells in the triangular either above or below the main diagonal. This question addresses issues such as trends in disagreement. For example, one rater may systematically select higher (or lower) scores, on ordinal rating scales. The third question concerns selections of disagreement cells, for example, the cells that are adjacent to the agreement cells in the diagonal. These cells indicate disagreement by just one scale point. The second and third questions are meaningful in particular when rating scales are used that are at least ordinal in nature. They can also be meaningful for nominal level ratings, for example, when cells with similar implications (in terms of, e.g., costs and treatment options) can be grouped into such patterns.

Is There Significant Lack of Disagreement?

For the examination of disagreement cells in a square rater $\times$ rater table, we propose derivatives of the coefficients κ, κ_n, and *ra.*

Cohen's κ as a Measure of Disagreement.

To introduce Cohen's κ as a measure of disagreement, we proceed as for the derivation of κ. First, we define the probability of disagreement as

$$\theta_1^d = \sum_{ij} p_{ij} = 1 - \sum_i p_{ii} = 1 - \theta_1.$$

Accordingly, we define the expected probability of disagreement under the assumption of rater independence as

$$\theta_2^d = \sum_{ij} p_{i.}\, p_{.j} = 1 - \sum_{i=j} p_{i.}\, p_{j} = 1 - \theta_2.$$

Using these measures, we can define the proportionate reduction in error measure of disagreement

Table 1. Cross-classification of facility and research diagnoses for 223 psychiatric patients (Fenig *et al.*, 1994); estimated expected cell frequencies in italic.

Facility Diagnosis	Research Diagnosis								
	1		2		3		4		Totals
1	40	*18.95*	6	*13.41*	4	*11.08*	15	*21.57*	65
2	4	*10.20*	25	*7.22*	1	*5.96*	5	*11.61*	35
3	4	*10.49*	2	*7.43*	21	*6.14*	9	*11.95*	36
4	17	*25.36*	13	*17.95*	12	*14.83*	45	*28.87*	87
Totals	65		46		38		74		223

$$\kappa^d = \frac{\theta_1^d - \theta_2^d}{1 - \theta_2^d} = \frac{(1 - \theta_1) - (1 - \theta_2)}{1 - (1 - \theta_2)} = \frac{-\theta_1 + \theta_2}{\theta_2}.$$

The coefficient κ^d indicates the reduction in the proportion of disagreements that results from comparing the observed with the expected frequencies in the off-diagonal cells, when the latter are estimated based on the assumption of rater independence. In the typical data situation in which disagreement is less likely than expected, we find that $\theta_2 < \theta_1$. Therefore, the numerator of the formula for κ^d is typically negative. Thus, the coefficient κ^d is typically negative, indicating that the agreement table contains fewer cases of disagreement than expected.

The range of κ^d is $-\frac{\theta_2^d}{1 - \theta_2^d} \leq \kappa^d \leq +1$. The upper limit is reached for $\theta_1 = 0$. The lower limit is reached for $\theta_1 = 1$. Thus, κ^d has a range comparable to κ. In fact, κ^d is the same measure as κ, just applied to the disagreement cells, and $\kappa^d = -\frac{\kappa(1 - \theta_2)}{\theta_2}$.

Brennan and Prediger's κ_n as a Measure of Disagreement

In a fashion similar to the one in the previous section, we can derive a Brennan and Prediger measure of disagreement, κ_n^d, as follows. The term θ_1^d is defined as for κ^d. The comparison term is derived from the base model that proposes a uniform probability distribution. In the off-diagonals, we find $2\binom{I}{2} = \frac{2I(I-1)}{2} = I(I-1)$ cells. Each of these has a probability of $1/I^2$. The total probability for all off-diagonal cells is thus $p_{ij} = \frac{I(I-1)}{I^2}$. We now set $\theta_{n,2}^d = p_{i \neq j}$, for the comparison term for the Brennan and Prediger coefficient of disagreement, and obtain

$$\kappa_n^d = \frac{\theta_1^d - \theta_{n,2}^d}{1 - \theta_{n,2}^d} = \frac{(1 - \theta_1) - \left(1 - \frac{1}{I}\right)}{1 - \left(1 - \frac{1}{I}\right)} = \frac{\frac{1}{I} - \theta_1}{\frac{1}{I}}.$$

As in the case of κ^d, we expect here that, in the usual data situation, in which raters tend to agree more than disagree, $\theta_1^d < \theta_{n,2}^d$. The coefficient will thus be negative, in most instances. The range of κ_n^d is $-\frac{\theta_{n,2}^d}{1 - \theta_{n,2}^d} \leq \kappa_n^d \leq 1$. As for κ and κ^d, κ_n^d is the same measure as κ_n, just applied to the disagreement cells. In fact, $\kappa_n^d = -\frac{\kappa_n(1 - 1/I)}{1/I} = -\kappa_n(I - 1)$.

The Coefficient *ra* as a Measure of Raw Disagreement.

In a fashion analogous to the previous two sections, the measure θ_1^d can be defined as a coefficient of raw disagreement, or ra^d. It is the complement of θ_1, because $\theta_1 + \theta_1^d = 1$. ra^d ranges $0 \leq ra^d \leq 1$.

The previous sections introduced the complements to three coefficients of agreement, the coefficients of disagreement. Barring differences in power that may result from differences in the number of agreements versus disagreements, null hypothesis tests concerning disagreement carry the same information as null hypothesis tests of agreement. The coefficients of disagreement carry no additional information, except that they express degree of disagreement in a form comparable to their agreement equivalents. In the following sections, the coefficients introduced above are used as starting points for the development of coefficients that address more specific questions concerning the structure of disagreement.

Disagreement in a Triangle of Off-Diagonal Cells

In the present section, we ask new questions. Specifically, we ask whether a specific pattern of disagreement exists. We focus on the upper triangle of an agreement matrix. This selection is arbitrary. Focusing on the lower triangle would lead to the same appraisals. We ask whether disagreement differs from chance. First, we define the chance model as the model of rater independence, as for Cohen's κ. Second, we define the chance model as the null model, as for Brennan and Prediger's κ_n. Third, we present the coefficient of raw disagreement for the upper triangle of an agreement table, analogous to the *ra* coefficients in the previous sections.

A coefficient of disagreement in the upper triangle of an agreement table that is analogous to Cohen's κ can be defined as follows. The probability of disagreement in those cells in which Rater B (columns) selects higher scores than Rater A (rows) is

$$\theta_1^t = \sum_{i<j} p_{ij} = 1 - \sum_i p_{ii} - \sum_{i>j} p_{ij},$$

where the superscript indicates that we consider the disagreement cells in the upper triangle. Accordingly, the expected probability of disagreements in the upper triangle is, under the model of rater independence,

$$\theta_2^t = \sum_{i<j} p_{i\cdot}\, p_{\cdot j} = 1 - \sum_i p_{i\cdot}\, p_{\cdot i} - \sum_{i>j} p_{i\cdot}\, p_{\cdot j},$$

and the resulting coefficient, κ^t, is

$$\kappa^t = \frac{\theta_1^t - \theta_2^t}{1 - \theta_2^t},$$

with a range of $-\infty < \kappa^t \leq 1$. The smallest possible value of κ^t is

$$\kappa_{\min}^t = \frac{-\theta_{2,\max}^t}{1 - \theta_{2,\max}^t} \text{ with } \theta_{2,\max}^t = \frac{(N - I - 1)^2}{N} + 2(I - 1)\frac{N - I - 1}{N} + \frac{\binom{I}{2} - 2(I - 1) - 1}{N}.$$

It should be noted that if $\theta_1^t = 1$, κ^t cannot be calculated based on the standard log-linear model of rater independence because $\theta_1^t = 1$ implies that the first column and the last row of the agreement table are empty. If this is the case, θ_2 is not always defined if the standard log-frequency model of rater agreement is used (for an illustration, see Appendix A). In this situation, two options may be considered. First, if the expected probabilities are not estimated from the data at hand but are derived from other information, a score for θ_2 can be determined. Second, if such information is unavailable, the column that contains Cell 11 and the row that contains Cell *II* can be declared structural zeros, and models of quasi-independence can be used to estimate expected cell frequencies. The resulting table can then be treated as other incomplete tables (Bishop, Fienberg, & Holland, 1975).

A Brennan and Prediger measure κ_n^t can be derived in an analogous manner. Specifically, θ_1^t is defined as for κ^t, and

$$\theta_{n,2}^t = \binom{I}{2}\frac{1}{I^2} = \frac{I - 1}{2I}.$$

We thus obtain

$$\kappa_n^t = \frac{\theta_1^t - \theta_{n,2}^t}{1 - \theta_{n,2}^t}.$$

The range of κ_n^t is $-\infty < \kappa_n^t < 1$. The smallest possible value of κ_n^t is

$$k_{n,\min}^t = \frac{-\theta_{n,2}^t}{I - \theta_{n,2}^t} = \frac{-(I - 1)/2I}{1 - (I - 1)/2I} = \frac{-I + 1}{I + 1}.$$

Following the same approach as before, the measure of raw disagreement is $ra^t = \theta_1^t$ with a range of $0 \leq ra^t \leq 1$.

The measures κ^t, κ_n^t, and ra^t assess disagreement in the upper triangular of the disagreement cells in an agreement table. κ^t compares the rate of disagreement to the rate that is expected under the model of rater independence. κ_n^t compares the rate of disagreement to the rate that is expected under the null model. ra^t indicates the proportion of incidences of disagreement in the selected cells. A trend exists if the measures κ^t, κ_n^t, and ra^t differ from the ones that result for all disagreement cells, that is, from the measures introduced in Section 3.1. Specifically, Rater B (columns) tends toward using higher scores on the ordinal rating scale than Rater A (rows) if $\kappa^d > \kappa^t$, $\kappa_n^d > \kappa_n^t$, and $ra^d > ra^t$. This applies accordingly for the comparison of the estimates for the upper with the estimates for the lower triangular of the agreement table, and to the case in which Rater B tends toward using lower scores than Rater A.

The problem addressed with κ^t is that one rater may tend to systematically assign higher rating categories. One might be tempted to use tests of differences in central tendency instead of κ^t. Clearly, such tests will allow one to test the null hypotheses that, on average, two or more raters use the same category level. However, these tests do not allow one to test this hypothesis under different base models. In addition, results from such tests do not correspond directly with results obtained with plain κ or its derivatives. Therefore, researchers may find κ^t useful because it does have the same interpretation as plain κ, and results are directly comparable.

Disagreement by One Scale Point

Thus far, the development of measures of disagreement has rarely used the idea that disagreement by one scale point may be less damaging than disagreement by more than one scale point (Cohen, 1968; Lawlis & Lu, 1972). Here, we focus on a selection of disagreement cells and ask whether, in those cells that contain the cases of disagreement by one scale point, disagreement differs from what was expected based on the chance models used for Cohen's κ and for Brennan and Prediger's κ_n. We develop measures in a fashion parallel to the measures in the previous sections.

A measure parallel to κ can be developed as follows. We define

$$\theta_1^1 = \sum_{i=1}^{I-1} p_{i,i+1} + \sum_{i=2}^{I} p_{i,i-1},$$

and, under the model of rater independence,

$$\theta_2^1 = \sum_{i=1}^{I-1} p_{i.}\, p_{.i+1} + \sum_{i=2}^{I} p_{i.}\, p_{.i-1}.$$

Together, these two measures can be used to define the κ-analogous measure of degree of disagreement by one scale point, $\kappa^1 = \dfrac{\theta_1^1 - \theta_2^1}{1 - \theta_2^1}$. The range of this measure is $\dfrac{-\theta_2^1}{1 - \theta_2^1} \leq \kappa^1 \leq 1$.

In an analogous way, we define a Brennan and Prediger–type measure of disagreement by one scale point using θ_1^1 and

$$\theta_{n,2}^1 = \frac{2(I-1)}{I^2},$$

and obtain

$$\kappa_n^1 = \frac{\theta_1^1 - \theta_{n,2}^1}{1 - \theta_{n,2}^1}.$$

The range of this measure is $\dfrac{-\theta_{n,2}^1}{1 - \theta_{n,2}^1} \leq \kappa_n^1 \leq 1$.

Finally, the coefficient of raw disagreement by one scale point is $ra^1 = \theta_1^1$, with a range of $0 \leq ra^1 \leq 1$.

The coefficients of disagreement by one scale point are of interest because researchers often consider disagreements by one scale point minor. Minor disagreements are expected to occur more often than disagreements by two or more scale points. Therefore, one would anticipate that null hypotheses concerning minor disagreements can be retained quite frequently.

In addition, the coefficients of disagreement by one scale point can be used to examine trends such that minor disagreement occurs more often than disagreement by two or more scale points. Specifically, if κ^1, κ_n^1, and ra^1 are closer to zero than their counterparts for all or the more extreme disagreement cells, one can conclude that disagreement by two or more scale points occurs less often, relative to expectancy, than disagreement by one scale point.

There have been early attempts to consider the case of disagreement by one scale point. Examples of such attempts include Cohen's weighted κ (see below in the discussion section), and Lawlis and Lu's (1972) coefficient. The relationship of the present approach to weighted κ is illustrated in the discussion section. The difference to Lawlis and Lu's coefficient is that κ^1, and κ_n^1 were designed to examine only the cells above and below the main diagonal, under different chance models, whereas Lawlis and Lu's coefficient was designed to examine agreement and one-point disagreement simultaneously. A version of κ^1 parallel to Lawlis and Lu's coefficient can be devised using the methods employed in this article.

Characteristics of the Coefficients of Disagreement: Simulation Results

In this section, we present simulation results concerning the distributional characteristics of the coefficients of disagreement presented in the preceding sections.

If κ and κ^d are complements of each other, their sampling distribution should be the same. The same applies to κ_n and ra and their counterparts, κ_n^d and ra^d. To illustrate this and, more generally, to describe the characteristics of the coefficients discussed in this article, a series of simulation runs were performed as follows. Square tables were created of sizes 3×3 through 10×10. For each of the tables, frequencies were randomly assigned to each cell. Two programs were written for the simulations. The first program was written in FORTRAN90. The random numbers were created using the generator RANDOM_NUMBER that is available in the MS Fortran Power Station. The generator returns uniformly distributed pseudorandom numbers within the interval $0 \leq m < 1.0$. This generator can be used to assign random numbers to arrays. The algorithm used to create the random numbers is that of a prime modulus M multiplicative congruential generator (Park & Miller, 1988). The resulting numbers were, for the present purposes, multiplied by 100 and then rounded to the nearest integer. The maximum cell frequency was set to a frequency that varied from $m_{ij,\max} = 5$ to $m_{ij,\max} = 50$. If a frequency ended up outside the predetermined range, it was replaced by a new random number that was subjected to the same procedure. Assignment of frequencies to cells was performed using multinomial sampling. For each maximum cell size, 1,000, 2,000, 3,000, or 5,000 samples were processed, thus varying both the number of samples and the maximum cell size. For each of the resulting sample distributions, the derivatives κ, κ_n, and *ra,* and the z statistic of κ, applied to the κ derivatives, were calculated.

To ensure reliability of results and to check whether the following simulation results are specific to the random number generator RANDOM_NUMBER, all runs were repeated using a second program written in the Mathematica environment (Wolfram, 1991). With the rounding algorithm and the random number generator the only exceptions, the programs were equivalent. Specifically, the random number generator in Mathematica, Random[], returns pseudorandom numbers on the interval $0 \leq m \leq 1.0$. As previously mentioned, this generator is used to assign values to an array. These random numbers were scaled to lie within the predetermined range of frequencies, and then rounded to the nearest

integer. This procedure guaranteed that replacement as described previously was not necessary. The parameters were varied and the coefficients of the resulting distributions were calculated as in the FORTRAN program.

The results obtained with the two simulations were virtually identical. Therefore, we use the results from the Mathematica program in all sections.

Characteristics of κ^d

We ask first, whether the measure κ^d and the corresponding z statistic are normally distributed. To answer this question, we create a probability plot that allows one to compare the observed distribution with the expected normal distribution. Figure 1 displays the distributions for for the maximum cell sizes of 10 and 50, runs with 1,000 samples, in tables of size 3×3.

The graphs in Figure 1 suggest that κ^d is normally distributed when the maximum cell size is 50 (see right panel; sample sizes varied from 91 to 387 with a mean of 224.98). For a maximum cell size of 10, κ seems to be slightly overdispersed and skewed (see left panel; sample sizes varied from 19 to 72 with a mean of 44.59). The variation of table sizes had the effect that the range of κ^d changed systematically. Specifically, the range of κ^d was largest for 3×3 tables, and smallest for 10×10 tables. To give an example, in 3×3 tables, κ^d ranged from -2 to $+1$, in 5×5 tables, κ^d ranged from -1.2 to 1, and for 7×7 tables, κ^d ranged from -1 to $+1$.

For significance testing, the distributions of the corresponding z statistics are more important than the distributions of the coefficients themselves. Therefore, we also examine the distributions of z for the two sample cases. The distributions of both the coefficients and their z statistics for the remaining samples, as well as for different cell sizes, show the same characteristics as the results shown here. Figure 2 shows the distributions of the z statistics for the cases shown in Figure 1.

The graphs in Figure 2 suggest that, for a maximum cell size of 10, the z statistic for κ is already nearly normally distributed. Its mean is close to 0, but it is, compared to the standard normal distribution, slightly skewed and underdispersed. In contrast, for a maximum cell size of 50, the statistic is better behaved. We conclude that this statistic can be used to test hypotheses concerning κ^d when maximum cell sizes are 10 or higher. However, when the maximum cell sizes are small, the binomial test may be preferable. The table size had no effect on the distribution of z_κ. The range of z_κ varied only with the maximum cell size.

The distributional characteristics of κ_n^d (not shown here), indicate that, for all cell sizes, the distributions are fairly symmetric, but slightly heavy tailed. In addition, the distributions show a wide range of κ_n^d values, including or approximating the extremes. The values of κ_n^d were also dependent on the size of the tables. Similar to κ^d, κ_n^d shows the largest range for small tables, and the smallest range for large tables.

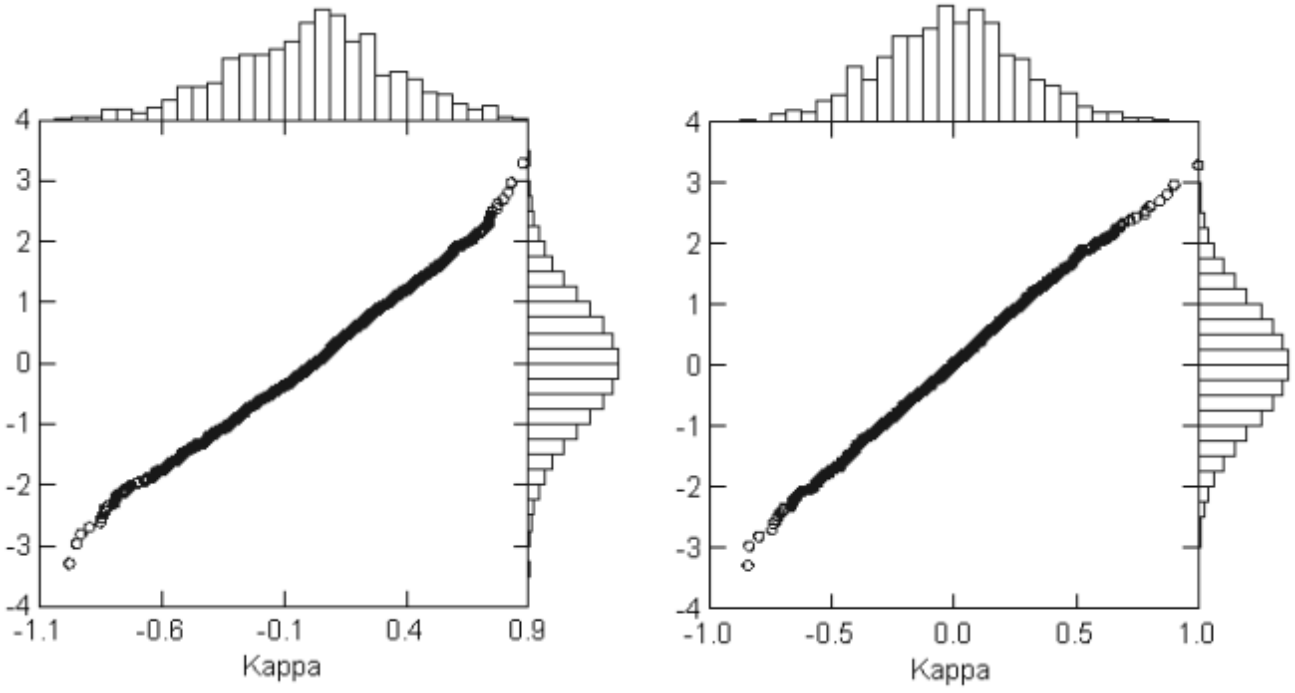

Figure 1. Probability plots for κ^d for the maximum cell sizes of 10 (left panel) and 50 (right panel), 1000 samples, in 3×3 tables.

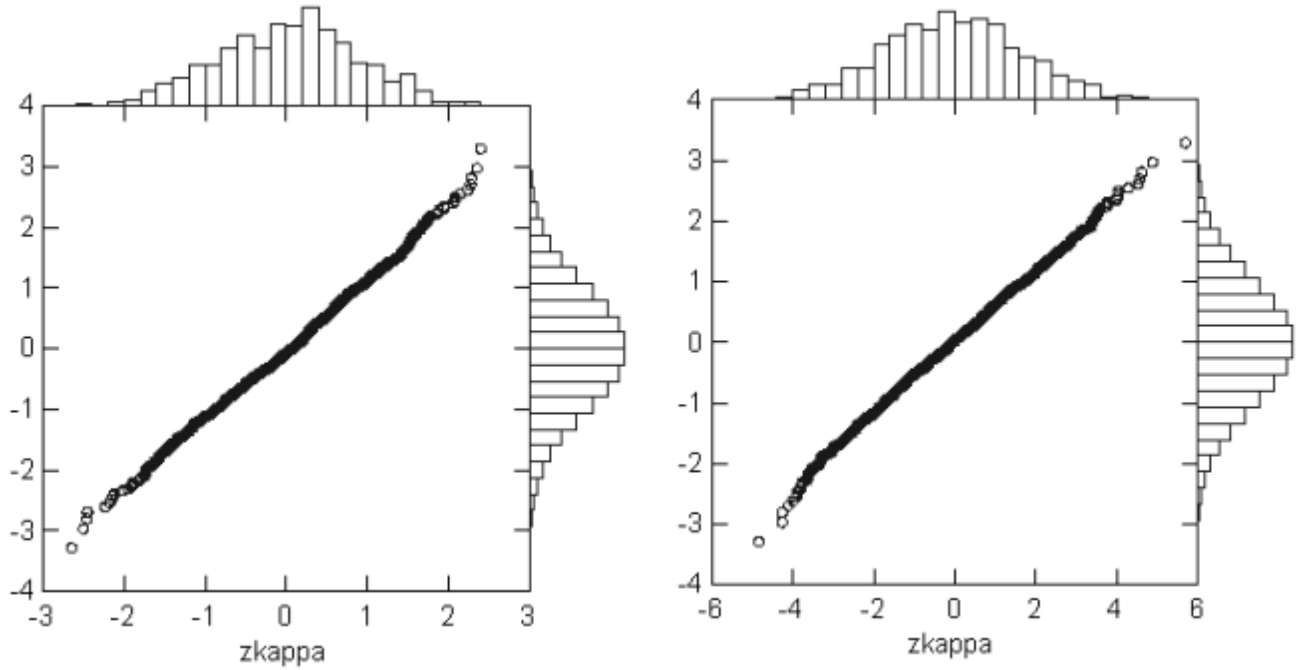

Figure 2. Probability plots for the z-statistics of κ^d for the maximum cell sizes of 10 (left panel) and 50 (right panel), 1000 samples, in 3 × 3 tables.

We now ask how the coefficient κ^d, the coefficient of raw agreement ra^d, and κ_n^d relate to each other. To answer this question, we examine the scatterplot matrix of these coefficients and the z statistic for κ in Figure 3. We use the data for the 1000 runs with a maximum cell size of 5, in 3 × 3 tables. The results for the larger maximum cell sizes, larger tables, and larger numbers of samples are, for all practical purposes, identical.

The scatterplot matrix shows first that the correlation between N and the measures included in this study is zero. In addition, each of the four measures is symmetrically distributed. The third result is that the correlations among the four measures are all very strong. The correlation between the measure of raw agreement and κ_n is equal to 1.0, as one would expect given that these two coefficients are linear transformations of each other. These results very closely replicate earlier results by von Eye (2005). However, the earlier results describe the behavior of these coefficients for the usual case in which agreement cells are examined; here, we study disagreement cells. In the next sections, we ask whether κ^d and the other measures can also be used for selections of disagreement cells.

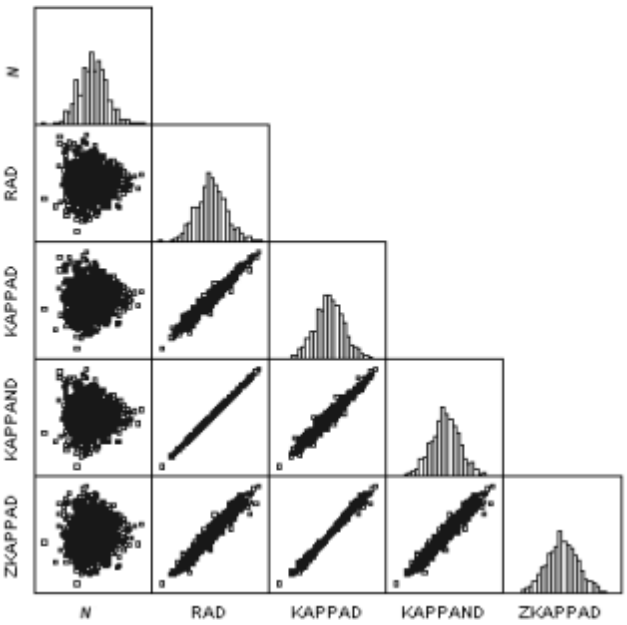

Figure 3. Scatterplot matrix of the coefficient of raw disagreement ra^d, κ^d, κ_n^d, and the z-statistic for κ^d, for a maximum cell size $m = 5$, 1000 samples, in 3 × 3 tables

Disagreement in a Triangular of Disagreement Cells: The Characteristics of κ^t, κ_n^t, and ra^t

In this section, we examine the distributions of the measures κ^t, κ_n^t, and ra^t as well as the z statistic for κ^t when applied to the upper triangular of an agreement matrix. The distribution of κ^t (not shown here) approximates the normal for all cell sizes. However, the range of κ^t is restricted (Figure 4). The reason for this restricted range is that the extreme cases in which all or almost all judg-

ments are in one of the triangulars did not occur in the simulations. They can occur under conditions other than those considered here (see Appendix A). Our simulations showed that both the number of runs and the size of the table influence the range of observed values for κ^t. For instance, a sample run with 10,000 samples increased the range of κ^t by 0.1 in both the positive and the negative domains, and the distribution of κ was close to perfectly normal. The dependency of κ^t on the size of the table was similar to the dependency found for κ^d. However, the largest range included only the values of -0.7 and $+0.9$ (for 3×3 tables), and shrunk to -0.2 to $+0.2$ for 7×7 tables.

One may wonder whether the distribution of the z statistic of κ^t can be expected to be restricted in range too. An inspection of the distribution of z (not shown here) indicates that the range of the statistic for κ is not restricted by the range of κ, but depends in its range on the maximum cell size. The statistic is fairly normally distributed, even for cells with maximum frequencies of 5, and, for cell sizes of up to 30, slightly heavy in its tails. We thus conclude that the statistic can be used to test hypotheses about κ^t, although the range of κ^t is restricted.

We also performed a comparison of κ^t to the measures κ_n^t, and ra^t. Results indicate first that the distributions of all calculated measures are symmetrical. None of the distributions is skewed, and only the distribution for the z-statistic for κ shows some elevated level of kurtosis. This can be explained as previously mentioned. Second, the comparison shows again that the correlation between κ_n^t and ra^t is 1.0. This is by definition. The correlations among the calculated measures are relatively high throughout, but lower than in the first simulation. Again, the restricted range is the reason. The lowest correlation is $r = 0.559$ between ra^t and κ^t. The highest is $r = 0.985$, between κ^t and z_κ. As previously, the correlations of the calculated measures with the sample size are all zero.

Disagreement by One Scale Point: The Characteristics of κ^1, κ_n^1, and ra^1

In this section, we examine the distributions of the measures κ^1, κ_n^1, and ra^1 as well as the z-statistic for κ^t when applied to the diagonals right above and right below the main diagonal of an agreement matrix. The distributions of κ^1 for tables of all simulated sizes and for all simulated cell sizes are very similar to each other, and not shown here. The only systematic difference was that the range of the κs was reduced for larger tables. This was parallel to the observation discussed in the previous sections.

The scatterplot matrix of the measures κ^1, κ_n^1, and ra^1 as well as the z-statistic for κ^1 are not shown here. That matrix indicates that the variable relationships are very similar to what we found in the previous sections. With increasing maximum cell sizes, numbers of samples, and table sizes, however, the distributions approximate those obtained for all off-diagonal cells in Figure 3.

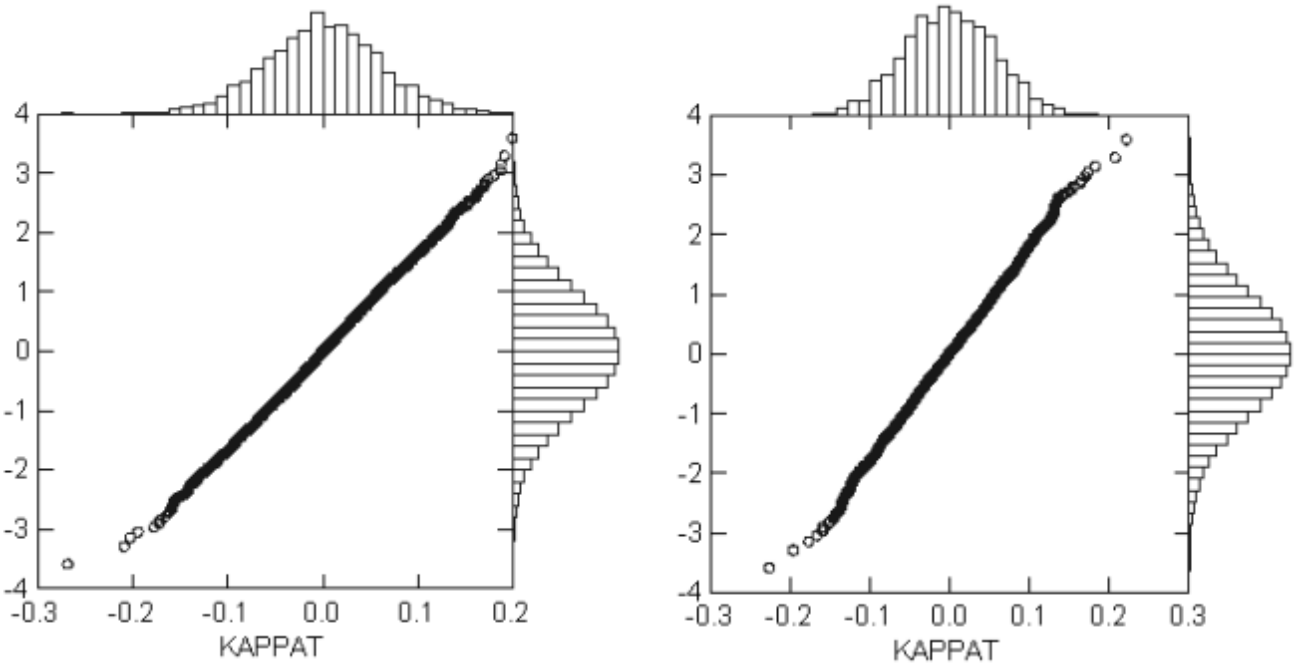

Figure 4. Probability plots for κ^1 for the maximum cell sizes of 10 (left panel) and 20 (right panel), for 3000 samples in 3×3 tables.

Data Example

In this section, we present a data example in which we analyze the data in Table 1 under the three scenarios discussed in this article. The estimated expected frequencies under the uniform distribution model that is used for Brennan and Prediger's (1981) κ_n, are all 223/16 = 13.9375.

For the case in which all disagreement cells are taken into account, we calculate $\kappa^d = (-0.5874 + 0.2743)/0.2743 = -1.142$. This value indicates that the difference between observed and expected frequencies in the disagreement cells is 0.5874 − 0.2743 = 0.3131. That is, the disagreement cells contain 31.31 percent fewer cases than expected under the assumption of rater independence. Weighted by the difference to the maximum number of cases that could possibly be found in the disagreement cells, we note that the reduction in error on the disagreement side is better than 114%.

Accordingly, we calculate $\kappa_n^d = -0.45(4 - 1) = -1.35$. Under the base model of a uniform distribution of judgments, the reduction in error is thus 135 percent. The coefficient of raw disagreement for all disagreement cells is calculated to be $ra^d = 1 - 131/223 = 1 - 0.5874 = 0.4125$. This indicates that 41.25% of the two institutions' judgments fail to be exact matches.

We now ask whether disagreement is particularly rare for those cases in which the Facility Diagnosis suggests less severe caseness than the Research Diagnosis. These are the cases in the upper triangle of off-diagonal cells. We calculate $\kappa^t = -0.24$. This value suggests that the proportionate reduction in these error cells is 24.14%, a value clearly less extreme than the one calculated for the entire table. For the κ_n equivalent, we calculate $\kappa_n^t = -0.31$, which indicates that 31.30% discrepant diagnoses are made less than expected under the assumption of a uniform diagnoses distribution. The coefficient of raw disagreement for the upper triangle of the disagreement cells is $ra^t = 0.18$.

Finally, we ask whether disagreement about the diagonal is less rare than disagreement by more than one scale point. First, we calculate $\kappa^1 = -0.19$, $\kappa_n^1 = -0.36$, and $ra^1 = 0.15$, indicating that 15.25% of the diagnoses are located in the cells one step away from the main diagonal. However, it can be hazardous to compare κ^t with κ^1 and κ^d, and, accordingly, the ra or the κ_n derivatives with each other. There are three problems that can prevent such comparisons from being valid. First, the range of the derivatives of κ depends on the number of summands used in the equation (compare Figures 1 and 4). Second, counterbalancing this effect only slightly, κ becomes more extreme when certain cells cannot be used for analysis. This comes into play when, for example, a coefficient κ^3 is calculated, that is, a κ-equivalent for disagreement three categories away from the main diagonal. In the present data example, this equivalent would prevent the middle two rows and the middle two columns from making a contribution to κ^3. Third, and most important, the reference used for each of the coefficients, θ_2, varies across the coefficients.

So, what can be done to answer the previous question? We propose two solutions. The first involves calculating the average raw residual. In our data example, we obtain for κ^d: −5.82; for κ^1: −4.96; for κ^2: −6.28; and, for κ^3: −7.47.[1] Looking at the values for κ^1, κ^2, and κ^3, it becomes clear that disagreement increases with the distance from the main diagonal.

The second solution involves taking the number of summands, n_s, into account. As the number of summands decreases, the value of θ_2 in the equation will decrease also, and so will the value of θ_1. However, the discrepancy in the denominator of $\kappa = \frac{\theta_1 - \theta_2}{1 - \theta_2}$ increases, and the value of κ decreases for given values of the numerator. One way to take the number of summands into account involves dividing the calculated κ values by the number of summands. This yields, for the present data, the values $\kappa^1/(n_s = 6) = -0.031$, $\kappa^2/(n_s = 4) = -0.035$, and $\kappa^3/(n_s = 2) = -0.04$, where n_s is given by the number of cells included in an analysis. In other words, the average negative reduction in error increases as we move away from the main diagonal. This result again supports the conclusion that disagreement increases as one moves away from the main diagonal.

Discussion

We begin the discussion with a summary of the characteristics of the 12 coefficients discussed here. Table 2 presents information on range, the case of agreement as expected from the base model, distribution and test statistics, and the selection of cells to which a coefficient is applied.

It was the goal of this research to study the behavior of three popular measures of rater agreement when applied to specific hypotheses concerning rater disagreement. Coefficients were derived for these applications, and simulations were performed to illustrate the behavior of these coefficients. Results showed that all coefficients and the *z*-statistic for the κ equivalents per-

1 κ^2 and κ^3 can be constructed parallel to κ^1. These measures help appraise disagreement two and three categories away from the main diagonal.

Table 2. Characteristics of 12 coefficients of rater agreement/disagreement.

Coefficient	Applied To	Range	No Effect If	Significance Tests
		Analysis of Agreement		
κ	diagonal cells	$-\theta_2/(1-\theta_2) \leq \kappa \leq 1$	$\kappa = 0$	z-, binomial test
κ_n	diagonal cells	$-\frac{1/I}{1-1/I} \leq \kappa_n \leq 1$	$\kappa_n = 0$	binomial test (von Eye & Mair, 2005)
ra	diagonal cells	$0 \leq ra \leq 1$	$ra = N/I$	binomial test
		Analysis of Disagreement		
κ^d	off-diagonal cells	$-\frac{\theta_2^d}{1-\theta_2^d} \leq \kappa^d \leq +1$	$\kappa^d = 0$	z-test
κ_n^d	off-diagonal cells	$-\frac{\theta_{n,2}^d}{1-\theta_{n,2}^d} \leq \kappa_n^d \leq 1$	$\kappa_n^d = 0$	binomial test
ra^d	off-diagonal cells	$0 \leq ra^d \leq 1$	$ra^d = N/(I(I-1))$	binomial test
κ^t	upper or lower triangular	$\frac{-\theta_{2,\max}^t}{1-\theta_{2,\max}^t} \leq \kappa^t \leq 1$	$\kappa^t = 0$	z-test
κ_n^t	upper or lower triangular	$(-I+1)/(I+1) \leq \kappa_n^t \leq 1$	κ_n^t	binomial test
ra^t	upper or lower triangular	$0 \leq ra^t \leq 1$	$ra^t = N/(0.5(I(I-1)))$	binomial test
κ^1	deviation from $i = j$ by ± 1	$\frac{-\theta_2^1}{1-\theta_2^1} \leq \kappa^1 \leq 1$	$\kappa^1 = 0$	z-test
κ_n^1	deviation from $i = j$ by ± 1	$\frac{-\theta_{n,2}^1}{1-\theta_{n,2}^1} \leq \kappa_n^1 \leq 1$	$\kappa_n^1 = 0$	binomial test
ra^1	deviation from $i = j$ by ± 1	$0 \leq ra^1 \leq 1$	$ra^1 = N/(2(I-1))$	binomial test

formed well under all conditions. Specifically, the coefficients performed well when all disagreement cells were taken into account, and when disagreement by only one scale point was considered. In the case of disagreement of the cells above the main diagonal of an agreement table, the coefficients performed differently. The reason for this difference is a restricted range of possible scores. When only the cells above (or below) the main diagonal are taken into account, the coefficients cannot reach the maximum score of 1.0 because the table under study will be incomplete. This restriction goes above and beyond the known restrictions for κ that result from unequal marginals.

One might consider using the more general models of quasi-independence for the estimation of the expected cell frequencies when hypotheses about the upper (or lower) triangular are tested. This option, however, is available only for tables of size 4×4 or greater. For smaller tables, the degrees of freedom become negative (see Appendix A).

It is interesting to note that the new coefficients, κ^d, κ^t, and κ^1 can be recast as special cases of Cohen's weighted κ. One benefit from using weighted κ is that specific aspects of the cross-classification of ratings can be examined. This article presents three examples of cases in which specific aspects of such cross-classifications are examined. However, instead of requiring that the user specify appropriate weight matrices, the present approach enables the user to test the same hypotheses using the simple arsenal provided by plain κ, using the coefficients proposed in this article. Thus, the specification of weight matrices is not required. It is easy to show that similar reformulations can be performed that allow one to inspect other selections of cells.

To illustrate, consider the two sets of cells, H and E. H contains the cells of interest, for example the disagreement cells, and E contains all other cells. Consider also the parameter ω, defined as

$$\omega_{ij} = \begin{cases} 1 \text{ if } ij \in H \\ 0 \text{ if } ij \in E. \end{cases}$$

Using ω, one can redefine θ_1 and θ_2 in a fashion analogous to the θ parameters for weighted κ, in Section 2.2. Using this notation, one can demonstrate the relationship of the measures proposed here to weighted κ, and recast the three cases using the following specifications for ω.

Case 1: All disagreement cells.

$$\omega_{ij} = \begin{cases} 1 \text{ if } i \neq j \\ 0 \text{ else.} \end{cases}$$

Case 2: Cells above main diagonal.

$$\omega_{ij} = \begin{cases} 1 \text{ if } i < j \\ 0 \text{ else.} \end{cases}$$

Case 3: Cells right above or below main diagonal.

$$\omega_{ij} = \begin{cases} 1 \text{ if } i = j - 1 \text{ or } i = j + 1 \\ 0 \text{ else.} \end{cases}$$

Accordingly, additional scenarios can be specified.

We conclude that the derivatives of κ can be used for the analysis of hypotheses concerning rater disagreement. The magnitude of the new coefficients is comparable with the magnitude of the original κ, even if the range of a coefficient is restricted under the described conditions. In all cases, the κ derivatives are interpretable as PRE measures. The significance statistics can also be interpreted. They do not seem to display restricted range, and they seem to work appropriately even for cell sizes as small as 5. Because the κ derivatives are interpretable as, and comparable to the original Cohen's κ, neither the specification of other statistical models nor the comparison of results from different statistical models is required. Therefore, users will find it easy to apply the new coefficients and interpret results.

References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research, 1,* 201–218.

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

Bishop, Y.M.M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687–699.

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.

DeCarlo, L. T. (2002). A latent class extension to signal detection theory, with applications. *Multivariate Behavioral Research, 37,* 423–451.

DeCarlo, L. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42,* 53–76.

Fennig, S., Craig, T. J., Tananberg-Karant, M., & Bromet, E. J. (1994). Comparison of facility and research diagnoses in first-admission psychotic patients. *American Journal of Psychiatry, 151*, 1423–1429.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323–327.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Graham, P. (1995). Modelling covariate effects in observer agreement studies: The case of nominal scale agreement. *Statistics in Medicine, 14,* 299–310.

Guggenmoos-Holzmann, I. (1995). Modelling covariate effects in observer agreement studies: The case of nominal scale agreement (letter to the editor). *Statistics in Medicine, 14,* 2285–2286.

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

Hsu, L. M., & Field, R. (2003). Interrater agreement measures: Comments on Kappa_n, Cohen's Kappa, Scott's π, and Aickin's α. *Understanding Statistics, 2,* 205–219.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lawlis, G. F., & Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin, 78,* 17–20.

Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96–107.

Park, S. K., & Miller, K. W. (1988). Random number generators: Good ones are hard to find. *Cacm, 31,* 1192–1201.

Schuster, C. (2001). Kappa as a parameter of a symmetry model for rater agreement. *Journal of Educational and Behavioral Statistics, 26,* 331–342.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology, 55,* 289–303.

Schuster, C., & Smith, D. A. (2002). Indexing systematic rater agreement with a latent class model. *Psychological Methods, 7,* 384–395.

Schuster, C., & von Eye, A. (2001). Models for ordinal agreement data. *Biometrical Journal, 43,* 795–808.

Tanner, M. A., & Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association, 80,* 175–180.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association, 88,* 421–427.

von Eye, A. (2002). *Configural Frequency Analysis—Methods, models, and applications*. Mahwah, NJ: Lawrence Erlbaum.

von Eye, A. (2005). An alternative to Cohen's κ. *European Journal of Psychology*. (in press)

von Eye, A., & Brandtstädter, J. (1988). Application of prediction analysis to cross-classifications of ordinal data. *Biometrical Journal, 30,* 651–655.

von Eye, A., & Mair, P. (2004). Significance tests for the measure of raw agreement. Under editorial review.

von Eye, A., & Mun, E. Y. (2005). *Modeling rater agreement—Manifest variable approaches*. Mahwah, NJ: Lawrence Erlbaum.

von Eye, A., & Sörensen, S. (1991). Models of chance when measuring interrater agreement with kappa. *Biometrical Journal, 33,* 781–787.

Wolfe, E. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46,* 35–51.

Wolfram, S. (1991). Mathematica. *A system for doing mathematics by computer* (2nd ed.). Redwood City, CA: Addison-Wesley.

Appendix A

Estimating κ for Triangles of Disagreement Cells

In this appendix, we illustrate the problems that arise when estimating κ coefficients for triangles of disagreement cells. Specifically, we show that tables (1) are incomplete and (2) can have negative degrees of freedom.

Consider the $I \times I$ cross-classification in which only the cells in the triangle above the main diagonal are frequented. In this table, the column that contains Cell 1 1 is empty, and the row that contains Cell I I is also empty. Table A1 illustrates this scenario using a 3×3 table.

Table A1 shows that 3×3 agreement tables that contain cases only in the triangular above the diagonal include six empty cells (shaded). In general, the degrees of freedom for estimating such a table under the base model of rater independence are, when each empty cell is blanked out costing one *df,*

$$df = I^2 - 1 - 2(I - 1) - \binom{I}{2} - I.$$

This equation shows that the base model of rater independence cannot be employed for small tables because degrees of freedom can become negative. The degrees of freedom will be negative for $I \leq 4$. This illustration helps explain why the range of κ^t values in Section 4.3 was limited: The simulation included only those tables for which the base model of rater independence was applicable without taking into account structural zeros.

Table A1. Cross-classification with empty cells (shaded).

		Rater B Rating Categories		
		1	2	3
Rater A Rating Categories	1			
	2			
	3			

Address for correspondence

Alexander von Eye
Michigan State University
Department of Psychology
107 D Psychology Building
East Lansing, MI 48824-1116
USA
E-mail: voneye@msu.edu

Or

Maxine von Eye
University of Bath
Department of Mathematical Sciences
Bath BA2 7AY

United Kingdom

