

# On the Marginal Dependency of Cohen's $\kappa$

Alexander von Eye<sup>1</sup> and Maxine von Eye<sup>2</sup>

<sup>1</sup>Michigan State University, East Lansing, MI, USA, <sup>2</sup>University of Cambridge, UK

**Abstract.** Cohen's  $\kappa$  (kappa) is typically used as a measure of degree of rater agreement. It is often criticized because it is marginal-dependent. In this article, this characteristic is explained and illustrated in the context of (1) nonuniform marginal probability distributions, (2) odds ratios that remain constant while  $\kappa$  changes in the presence of varying marginal distributions, and (3) percentages of raw agreement that remain constant while  $\kappa$  changes in the presence of varying marginal distributions. The meaning and interpretation of  $\kappa$  are explained with reference to the log-linear main effect model of variable independence. This model is used for the estimation of the expected cell frequencies of agreement tables. It is shown that the interpretation of  $\kappa$  as a measure of degree of agreement is incorrect. The correct interpretation is that  $\kappa$  assesses the degree of agreement beyond that expected based on a statistical model such as the independence or the null model. Based on Goodman's (1991) distinction between marginal-free and marginal-dependent measures, it is shown that  $\kappa$  is marginal-dependent. It shares this characteristic with the well-known  $\chi^2$ -statistic and the correlation coefficient for cross-classifications. In contrast, the odds ratio, the unweighted log-linear interaction, and the percentage of raw agreement are marginal-free. Therefore, the expectation that marginal-dependent  $\kappa$  would reflect the same data characteristics as some of the marginal-free measures is misguided. It is recommended that researchers report both measures of degree of agreement and measures of agreement beyond some expectation.

**Keywords:** kappa, marginal dependency, marginal free, marginal dependent, chance model, odds ratio

Cohen's  $\kappa$  (kappa; 1960) is the most frequently used measure of rater agreement. It takes the probabilities with which raters use rating categories into account. Because of this characteristic,  $\kappa$  is known to be *marginal-dependent* (also called *prevalence-dependent*; see Agresti, 2002; Guggenmoos-Holzmann, 1995). In this article, we first review  $\kappa$ . We then describe, exemplify, and discuss the characteristic of marginal dependency. We also show that, under specific conditions, marginal dependency does not apply. We explain marginal dependency based on (1) the chance model that is used to estimate expected cell frequencies, and (2) Goodman's (1991) concept of marginal-free versus marginal-dependent measures. In a simulation, we compare  $\kappa$ , the odds ratio, and raw agreement. Finally, we make recommendations for researchers who wish to express the degree of rater agreement using easy-to-interpret coefficients.

## Cohen's $\kappa$

To introduce Cohen's (1960) kappa ( $\kappa$ ), consider the cross-tabulation of the judgments that two raters provide (see also Mun, 2005). These cross-tabulations are also called *agreement tables*. Consider the two raters, A and B, who used the three categories 1, 2, and 3 to evaluate students' performance in a test. The agreement table of these raters' judgments is given in Table 1 (see von Eye & Mun, 2005).

The interpretation of the frequencies,  $m_{ij}$ , in the cross-

Table 1. Agreement table of two raters' Judgments

		Rater B		
		Rating categories		
		1	2	3
Rater A	1	$m_{11}$	$m_{12}$	$m_{13}$
Rating Categories	2	$m_{21}$	$m_{22}$	$m_{23}$
	3	$m_{31}$	$m_{32}$	$m_{33}$

classification given in Table 1 is straightforward: Cell 1 1 displays the number of instances in which both Rater A and Rater B used Category 1; Cell 1 2 contains the number of instances in which Rater A used Category 1 and Rater B used Category 2, and so forth. The cells with indexes  $i = j$  with  $i, j = 1, \dots, 3$  display the numbers of incidences in which the two raters used the same category. These cells are also called the *agreement cells*. These cells are shaded in the table. All other cells can be called *disagreement cells*.

In the following paragraphs, we derive  $\kappa$  (see von Eye & Mun, 2005). Let  $p_{ij}$  be the probability of Cell  $ij$ , with  $i = 1, \dots, I$  and  $j = 1, \dots, J$ . The *agreement cells* have probability  $p_{ii}$ . The parameter  $\theta_1$ ,

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

describes the proportion of instances in which the two raters agree. To have a reference value with which to compare

$\theta_1$ ,  $\kappa$  assumes independence of the two raters. In other words, it is assumed that the raters do not influence each other when providing their judgments. In addition, each rater can use the rating categories as often as they please. Based on these assumptions, we can estimate the proportion of instances in which the two raters agree by chance using the reference value  $\theta_2$ ,

$$\theta_2 = \sum_{i=1}^I p_{i.} p_{.i}$$

where a period indicates the marginal summed across. More specifically,  $i$  indicates the  $i$ th row probability, and  $.i$  indicates the  $i$ th column probability. Subtracting  $\theta_2$  from  $\theta_1$  results in a measure of rater agreement that takes the assumption of rater independence into account. If the difference  $\theta_1 - \theta_2$  is positive, the two raters agree more often than expected based on this assumption. If  $\theta_1 - \theta_2$  is negative, they agree less often than expected.

The largest possible discrepancy between  $\theta_1$  and  $\theta_2$  is  $1 - \theta_2$ . This discrepancy results when all judgments appear in the agreement cells of the cross-classification (see Table 1). In this case, agreement is perfect. Weighting the difference  $\theta_1 - \theta_2$  by  $1 - \theta_2$  yields Cohen's  $\kappa$ ,

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

$\kappa$  indicates the proportion of incidences in which two raters use the same categories to evaluate a number of objects, in comparison to the expected proportion, relative to the maximally possible difference between the observed and the expected proportions.

It is important to realize that the model that is used to estimate the expected frequencies for Cohen's  $\kappa$  corresponds to the log-linear main effect model (Agresti, 2002). In other words, the chance model takes main effects into account which reflect the raters' differential use of the  $I$  rating categories, and it assumes independence between raters. This model can be cast as

$$\log m = \lambda + \lambda_i^A + \lambda_j^B,$$

where the subscripts indicate the main effect parameters, and the superscripts indicate the raters. The same model underlies Cohen's (1968) weighted  $\kappa$  which allows one to take weights of disagreements into account.

Alternative chance models have been discussed. For example, Brennan and Prediger (1981) proposed using the null model as a chance model, that is  $\log m = \lambda$  (for a comparison of Cohen's with Brennan and Prediger's  $\kappa$ , see von Eye & Sørensen, 1991; for a comparison with other manifest variable models, see von Eye & Mun, 2005). Each of these models allows the researcher to calculate statistics that are parallel to Cohen's  $\kappa$ .

One interpretation of  $\kappa$  can be based on the characteristic

of  $\kappa$  as a measure of *proportionate reduction in error* (PRE; Fleiss, 1975). Using  $\kappa$ , researchers inspect the cells in the main diagonal of the  $I \times I$  cross-classification of two raters' judgments. The question asked is how the observed frequency distribution differs from the expected, or chance, distribution in the diagonal cells. If the agreement cells contain more cases in the observed distribution, one expresses the result in terms of the proportionate reduction in error. This reduction indicates that the observed frequency distribution contains more cases in the agreement cells and fewer cases in the *disagreement cells* than the chance distribution. Specifically,  $\kappa$  is an example of a PRE measure of the form

$$PRE = \frac{\theta_1 - \theta_2}{\max(\theta_1) - \theta_2}$$

The maximum value that  $\theta_1$  can take is 1. This would indicate that all responses are located in the main diagonal, or that there are no disagreements. The above definition of  $\kappa$  uses  $\max(\theta_1) = 1$  in the denominator. Thus,  $\kappa$  can be identified as a PRE measure.

For the discussion in this article, one important argument is that the proportionate reduction in error is interpreted with respect to a particular chance model. That is,  $\kappa$  indicates the proportion of agreement cases that exceeds an expected proportion. Therefore, interpretation of  $\kappa$  as a measure of *degree of agreement* is incorrect. By implication,  $\kappa$  can be low even if the absolute number of agreement cases is large. For example, for  $m_{11} = 1$ ,  $m_{12} = 10$ ,  $m_{21} = 3$ , and  $m_{22} = 60$ ,  $ra = 0.82$ , indicating that 82% of the judgments are in complete agreement. For the same table,  $\kappa = 0.06$ , indicates weak agreement beyond the chance model. Therefore, the rules of thumb concerning the magnitude of  $\kappa$  that can be found in the literature are highly problematic. Consider the well-known classification by Landis and Koch (1977), according to which

- $\kappa < 0.00$             poor agreement
- $0.00 < \kappa \leq 0.20$    slight
- $0.20 < \kappa \leq 0.40$    fair
- $0.40 < \kappa \leq 0.60$    moderate
- $0.60 < \kappa \leq 0.80$    substantial
- $0.80 < \kappa \leq 1.00$    almost perfect agreement.

Clearly, this classification is based on an interpretation of  $\kappa$  as a measure of degree of agreement instead of the degree to which agreement exceeds expectancy. Similarly, statements such as "the stronger the agreement, the higher is  $\kappa$ , for given marginal distributions" (Agresti, 2002, p. 434) are problematic. In the section "Marginal Dependency of Cohen's  $\kappa$ " below, we discuss this issue in more detail, focusing on the interpretation of  $\kappa$  as a measure of agreement beyond expectation.

Under a multinomial sampling scheme, an estimator of  $\kappa$  is

$$\hat{\kappa} = \frac{N \sum_i m_{ii} - \sum_i m_i m_i}{N^2 - \sum_i m_i m_i},$$

where  $i = 1, \dots, I$  indexes the rating categories,  $N$  is the number of decisions made by the raters, and  $m$  indicates the observed frequencies. This measure has been used and discussed extensively (for overviews see, e.g., Agresti, 2002; Fleiss, Levin, & Paik, 2003; von Eye & Mun, 2005; Wickens, 1989).  $\kappa$  is strongly related to  $\lambda$ , a measure of (asymmetric) similarity (Froman & Llabre, 1985; Goodman, & Kruskal, 1954).

The characteristics of  $\kappa$  include

1. The range of  $\kappa$  is  $-\infty < \kappa \leq 1$ ; positive values of  $\kappa$  reflect agreement better than chance, and negative values of  $\kappa$  reflect agreement less than chance.
2.  $\kappa = 0$  if the probability of disagreement is the same as the probability of agreement;  $\kappa$  can be zero even if the raters' judgments are not independent.
3.  $\kappa = 1$  only if the probability of disagreement is zero.
4.  $\kappa$  is defined only if at least two categories are used by both raters, that is, if the probability,  $p_{ij}$ , is greater than zero for at least two cells.
5. If the probability in the off-diagonals is nonzero, the maximum value of  $\kappa$  decreases as the marginals deviate from a uniform distribution (see the notion of *prevalence dependency* of chance-corrected agreement; Cook & Farewell, 1995; Guggenmoos-Holzmann, 1995). In the section "Marginal Dependency of Cohen's  $\kappa$ ," this characteristic is discussed in more detail.
6. When the probability of disagreement decreases and is smaller than the number of agreements,  $\kappa$  increases monotonically; when the probability of disagreement increases and is greater than the probability of agreement,  $\kappa$  does not decrease monotonically (von Eye & Sörensen, 1991).

## Data Example

The following example, taken from Bortz and Lienert (1998, p. 270), analyzes data from a study on adolescents with behavior problems. Two psychiatrists, A and B, classify 100 adolescents using the categories neglected (NEG), neurotic (NEU), and psychotic (PSY). We ask whether the psychiatrists agree better than chance. Table 2 displays the observed frequencies and, in the same cells, the estimated expected cell frequencies (*in italics*) that were estimated using the log-linear main effect of rater independence.

The Pearson  $\chi^2 = 38.59$  ( $df = 4$ ;  $p < .01$ ) suggests that the two psychiatrists are not independent in their ratings. We now ask whether above-chance agreement can be the reason for this. We estimate  $\kappa = 0.43$  (CI:  $0.26 \leq \kappa \leq 0.58$ ), which indicates that the psychiatrists agree significantly more than expected based on the main effect chance model.

Table 2. Agreement between two psychiatrists: Observed and expected frequencies (*in italics*)

		Psychiatrist B			totals
		NEG	NEU	PSY	
Psychiatrist A	NEG	53 <i>39</i>	5 <i>15</i>	2 <i>6</i>	60
	NEU	11 <i>19.5</i>	14 <i>7.5</i>	5 <i>3</i>	30
	PSY	1 <i>6.5</i>	6 <i>2.5</i>	3 <i>1</i>	10
totals		65	25	10	100

In the following sections, we explicate characteristics and interpretation of  $\kappa$ , with a focus on marginal dependency.

## Marginal Dependency of Cohen's $\kappa$

As Agresti noted (2002, p. 435), "controversy surrounds the utility of kappa . . . , partly because their values depend strongly on the marginal distributions." This characteristic of  $\kappa$  has at least three facets. The first is that  $\kappa$  cannot even approximate its maximum value of 1.0 when the marginal distributions are not uniform. The second is that  $\kappa$  can differ in magnitude even when the association between two raters' judgments is constant, but the marginal distributions vary. The third facet to be discussed here is that  $\kappa$  can vary even if the coefficient of raw agreement (which reflects percent agreement) is constant.

### Marginal Dependency of $\kappa$ for Extreme Marginal Distributions

The first facet of our discussion examines the purported inability of  $\kappa$  to reach its maximum value of 1.0 when the marginal distributions are not uniform. To demonstrate this characteristic, consider the two cases in Example 1, in Table 3.

In the left hand panel of Table 3 (Case 1), we find two-thirds of the judgments in the agreement cells. For  $\kappa$ , we calculate 0.40, which indicates that 40% more judgments than expected under the log-linear base model are located in the diagonal cells.

For the following considerations, we need the *coefficient of raw agreement*. It is defined as  $ra = \sum_i p_{ii}$ .

Table 3. Example 1 of the marginal dependency of  $\kappa$

Case 1	Case 2		
	1	2	Totals
1	1	1	2
2	0	1	1
Totals	1	2	3

Case 1	Case 2		
	1	2	Totals
1	1	10,000	1
2	0	1	1
Totals	10,000	2	10,002

Table 4. Example 2 of the marginal dependency of  $\kappa$ ; Odds ratio is 10 for each of the panels

Case 1	Case 2			Case 3							
	1	2	Totals	1	2	Totals	1	2	Totals		
1	3799	1201	5000	1	1930	1070	3000	1	412	588	1000
2	1201	3799	5000	2	1070	5930	7000	2	588	8412	9000
	5000	5000	10000	3000	7000	10000		1000	9000	10000	

In the right panel (Case 2), we find  $ra = 0.9999$  of the judgments in the agreement cells. Still,  $\kappa$  is no greater than 0.667. Increasing the frequency in Cell 11 to 1,000,000,000 will not give us a larger value of  $\kappa$ ; it will still be 0.667. If  $\kappa$  were a measure of degree of agreement, one would expect it to approach 1.0 as the percentage of judgments in agreement cells increases.

This phenomenon is known as marginal dependency (see Guggenmoos-Holzmann, 1995). However, the phenomenon of marginal dependency carries only so far. As soon as the  $ra = 0.9999$  is changed to 1.00,  $\kappa$  leaps to 1.0 – regardless of marginal distribution. This is illustrated in Figure 1. The top and the bottom lines indicate  $\kappa$  for the frequencies of 1, 10, 100, and 1000, all in Cell 1 1. The bottom line indicates  $\kappa$  for all of these frequencies, with the frequency in Cell 1 2 set to one, as in Table 3. The top line indicates  $\kappa$  for all of these frequencies, with the frequency in Cell 1 2 set to zero. Clearly, the characteristic of marginal dependency applies only when raw agreement is  $ra < 1.00$ . To the best of our knowledge, this has not been discussed in the literature.

The middle line in Figure 1 illustrates the behavior of

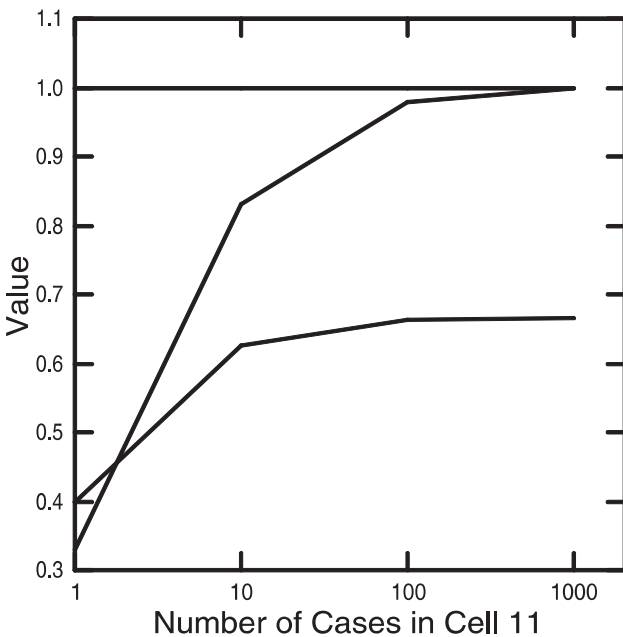


Figure 1. Marginal dependency of  $\kappa$  under conditions of perfect (top line) and less than perfect agreement (bottom line); the middle line illustrates Brennan and Prediger's (1981)  $\kappa$  under the less than perfect agreement condition.

Brennan and Prediger's (1981) version of  $\kappa$  under the less than perfect agreement condition. As was mentioned in the section "Cohen's  $\kappa$ ," this version of  $\kappa$  uses the null model for a reference. Therefore, the resulting version of  $\kappa$  is not margin-dependent.

The examples in Table 3 and Figure 1 illustrate that

1.  $\kappa$  will not approximate the maximum value of 1 when marginal frequencies are not uniform, as long as the probability for at least one disagreement cell is greater than zero (bottom line in Figure 1); this demonstrates the marginal dependency of  $\kappa$ ;
2.  $\kappa$  will assume the maximum value of 1 as soon as the probability for all disagreement cells is zero (middle line in Figure 1); thus,  $\kappa$  shows no marginal dependency under conditions of perfect agreement.

### Marginal Dependency of $\kappa$ for Constant Association Values

The second facet of marginal dependency for which  $\kappa$  is being criticized can be described as follows. The same judgment process, described by some measure of association, can result in different values of  $\kappa$ , depending on marginal distribution. In the examples used in Agresti (2002, p. 453, Example 10.40; see Cook, 1998), the "same judgment process" in  $2 \times 2$  tables was operationalized by an odds ratio that was constant in the presence of varying marginal distributions. Specifically, consider  $2 \times 2$  matrices whose cell probabilities can be found, given the marginal probabilities  $p_{1.} = p_{.1} = \beta$  and the odds ratio

$$\frac{p_{11} p_{22}}{p_{12} p_{21}} = \theta,$$

by  $p_{11} = \beta - p_{12}$  and  $p_{22} = (1 - \beta) - p_{12}$  (for more detail, see Appendix A). In Table 4, we illustrate the examples using the odds ratio  $\theta = 10$ , and  $\beta = 0.5, 0.3$ , and  $0.1$ . The following numerical example illustrates that  $\kappa$  can vary even when the odds ratio stays constant. The three panels in Table 4 contain frequencies that correspond to the odds ratio and marginal probabilities specified in Agresti's (2002) example 10.40.

The odds ratio in each of the three panels of Table 4 is 10.0. However, the corresponding values of  $\kappa$  are 0.52, 0.49, and 0.35, thus reflecting that  $\kappa$  is smaller for non-uniform marginal distributions than for uniform marginal distributions, given a constant odds ratio. Figure 2 illustrates the relationship between the marginal distribution

and  $\kappa$  for symmetric marginal distributions and constant odds ratio.

### Marginal Dependency of $\kappa$ for Constant Raw Agreement

Introduced in the section "Marginal Dependency of  $\kappa$  for Extreme Marginal Distributions," the coefficient of raw agreement is defined as  $ra = \sum_i p_{ii} = \theta_1$ . To demonstrate that  $\kappa$  can vary for constant values of  $ra$  (see von Eye, 2008), consider the example in Table 5. In this example, raw agreement in each of the three panels is 0.99. The  $\kappa$  values for the three panels are 0.98, 0.97, and 0.66, thus again reflecting the characteristic that its values depend on the marginal distribution.

Figure 2 depicts the behavior of  $\kappa$  when the odds ratio  $\theta$  and the coefficient  $ra$  are held constant. In each case,  $\kappa$  reaches a maximum when the marginal distribution is uniform, that is, when the ratio of marginal probabilities is 1.0 (or the ratio of the first marginal to the total is 0.5; see Figure 2). In general, it can be shown that  $\kappa$  is sensitive to data characteristics such as marginal probabilities that other measures are not sensitive to. The fact that  $\kappa$  is sensitive to characteristics of marginal distributions should not come as a surprise, as will be shown in the next section.

Table 5. Example 2 of the marginal dependency of  $\kappa$ ; raw agreement is 0.99 for each of the three panels

Case 1	Case 2			Case 3							
	1	2	Totals	1	2	Totals	1	2	Totals		
1	49	1	50	1	24	1	25	1	1	2	
2	0	50	50	2	0	75	75	2	0	98	98
	49	51	100	24	76	100		1	99	100	

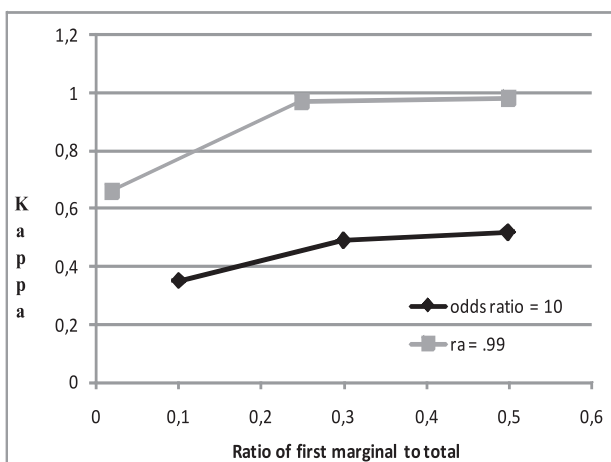


Figure 2. Relationship between ratio of marginals and  $\kappa$ , for symmetric distributions in  $2 \times 2$  tables and constant odds ratio  $\theta = 10$  (Table 4), and for constant  $ra = 0.99$  (Table 5).

In this next section, we explain the marginal dependency of  $\kappa$ .

### Explaining the Marginal Dependency of $\kappa$

To explain the behavior of  $\kappa$  under the conditions illustrated in the previous section, we first review the definition and estimation of  $\kappa$  and then look at the effects that the conditions described above have on the estimate.

For the definition of  $\kappa$  given above, a different computational approach can be taken than the one presented in the section "Cohen's  $\kappa$ ." Specifically, under the main effect model of rater independence, the expected probabilities for the agreement cells can be expressed in a way similar to the way used for the well-known  $\chi^2$  statistic. Thus, we obtain for the expected frequencies in the agreement cells

$$e_{ii} = \frac{m_{i.} m_{.i}}{N}$$

where  $m_{i.}$  indicates the row totals and  $m_{.i}$  the column totals of the agreement table. Again, in a fashion parallel to the well-known  $\chi^2$  statistic, we then obtain for Cohen's  $\kappa$  the estimator

$$\hat{\kappa} = \frac{\sum_i m_{ii} - \sum_i \frac{m_{i.} m_{.i}}{N}}{N - \sum_i \frac{m_{i.} m_{.i}}{N}}$$

This expression shows that

1. The sum of the observed frequencies in the agreement cells is compared to the sum of the expected frequencies in the agreement cells;
2. For the estimation of the expected cell frequencies the row totals and the column totals are taken into account. This reflects the fact that the main effect model of rater independence is used as a chance model; and
3.  $\kappa$  quantifies the weighted difference between the sums of the observed and the expected frequencies in the agreement cells (similar measures can be defined for selections of disagreement cells; see von Eye & von Eye, 2005).

The second of these characteristics suggests that different models than the main effect model can be considered as chance models for  $\kappa$  measures. Indeed, Brennan and Prediger (1981) proposed using the null model,  $\log m = \lambda$ . This model assumes a uniform marginal distribution for both raters and, as a consequence, is not marginal-dependent. A significance test for Brennan and Prediger's  $\kappa$  has been proposed by von Eye, Schauerhuber, and Mair (2007). Other models are conceivable, for instance, ordinal models or models with weights.

The second of these characteristics shows also that  $\kappa$  cannot be interpreted as a measure of *degree of agreement*. Instead,  $\kappa$  expresses the degree to which observed agreement *exceeds the agreement that was expected under some chance model*. This is the main reason why rules of thumb such as the one given by Landis and Koch (1977) are problematic. Marginal dependence is another reason.

Using these arguments, we can now proceed and show why  $\kappa$  is sensitive to the marginal distribution. We discuss the examples from the section "Marginal Dependency of Cohen's  $\kappa$ " in the same order.

### Explaining the Marginal Dependency of $\kappa$ for Extreme Marginal Distributions

To explain the behavior of  $\kappa$  when marginal distributions are extreme, consider the example in Table 3. In this example,  $\kappa$  was far from its maximum score of 1.0 even when raw agreement was 99.99%. Table 6 displays the expected cell frequencies for this example.

The comparison of the observed cell frequencies (Table 3) and the expected cell frequencies (Table 6) shows that, in the first case, when the marginals contain one-third and two-thirds of the sample, the deviation from expectancy is, in each cell, 0.33. That is, each agreement cell contains one-third more judgments than expected. Each of the disagreement cells contains one-third fewer cases than expected. Considering that the disagreement cells are not all empty, it makes sense that  $\kappa$  is less than 1. In the second case, we note that the expected frequencies are close to the observed frequencies again. Specifically, each of the agreement cells contains one judgment more than expected, and each of the disagreement cells contains one judgment fewer than expected. In other words, the log-linear base model is able to nicely reproduce the joint distribution with the uneven marginals. The proportion of cases in the agreement cells that exceeds expectation is, therefore, less than 100%. In the present example, we find that three judgments are

expected for the disagreement cells, but only one was found. We, thus, can say that only one-third of the expected judgments was found, which amounts to a discrepancy of two-thirds = 0.667, equaling the value of  $\kappa$ . In other words, in the second panel, we find only 66.7% fewer cases in the disagreement cells than expected. The chance model reproduces the observed distribution well. However, the number of disagreements was overestimated by two-thirds.

The implication of this example for the characteristics and interpretation of  $\kappa$  is clear. If  $\kappa$  is used as a measure of the degree of agreement as suggested by some authors, it will fail to indicate that agreement is strong when a marginal distribution is uneven. The magnitude of  $\kappa$  will be conditional on a marginal distribution, which can be unsatisfactory. However,  $\kappa$  is not a measure of the degree of agreement. Instead, it is a measure of agreement above and beyond expectation. As was indicated above, a number of chance models has been discussed, including the main effect model (Cohen, 1960) and the null model (Brennan & Prediger, 1981).  $\kappa$  indicates the proportion of judgments that are found in agreement cells above and beyond expectation. In this respect,  $\kappa$  performs, in the first example (and in the following ones), as designed.

### Explaining the Marginal Dependency of $\kappa$ for Constant Odds Ratios

To explain the behavior of  $\kappa$  when the odds ratio in tables is constant but the marginal distributions vary, consider again the example in Table 4. The expected cell frequencies for this example are given in Table 7.

The comparison of the observed frequencies (Table 4) with the corresponding expected frequencies (Table 7) shows that, in the first panel, each cell deviates from expectation by 1,299 judgments. In the agreement cells, this is the number by which expectation is exceeded, and in the disagreement cells, this is the number below expectation. In the disagreement cells, 48% of the expected judgments

Table 6. Expected cell frequencies for example 1 of the marginal dependency of  $\kappa$  (Table 3)

Case 1	Case 2			Case 2	Case 2		
	1	2	Totals		1	2	Totals
1	.67	1.33	2	1	9999	2	10001
2	.33	.67	1	2	1	0	1
Totals	1	2	3	Totals	10000	2	10002

Table 7. Expected cell frequencies for example 2 of the marginal dependency of  $\kappa$ ; the odds ratio is 10 for each of the three panels

Case 1	Case 2			Case 3							
	1	2	Totals	1	2	Totals	1	2	Totals		
1	2500	2500	5000	1	900	2100	3000	1	100	900	1000
2	2500	2500	5000	2	2100	4900	7000	2	900	8100	9000
	5000	5000	10000		3000	7000	10000		1000	9000	10000

are observed. In different words, the agreement cells contain 52% more judgments than expected. This is the value of  $\kappa$ .

For the same sample size and odds ratio, the deviation from expectancy in the second panel of Tables 4 and 7 is 1,030, for each cell. This discrepancy is smaller than the one found for the first panel. Accordingly, the value for  $\kappa$  is smaller for the second panels than for the first, too. The reason for this result is that the chance model is able to reproduce the frequency distribution in the second panel better than the one in the first panel. In the second panel, there is a main effect, whereas in the first panel, there is none.

In the third panel, the main effect is even stronger. The chance model, therefore, comes even closer to reproducing the frequency distribution, and the deviation in each cell is down to 312. Accordingly, the value of  $\kappa$  is the smallest in this panel (0.66 vs. 0.96 and 0.97 for the first two panels).

We, thus, conclude again that  $\kappa$  is a measure that describes agreement beyond expectation instead of amount or degree of agreement. When a chance model is able to describe a distribution well, the portion of judgments that is left for *agreement beyond chance* will, naturally, be small. This applies regardless of the magnitude of the odds ratio.

### Marginal-Free Versus Marginal-Dependent Measures

Another explanation for the different behavior of the odds ratio and  $\kappa$  can be given based on a classification that was proposed by Goodman (1991). The author begins with a look at the independence model given in the section "Cohen's  $\kappa$ ." This model proposes that  $p_{ij} = p_{i.} p_{.j}$ . It implies that nonindependence is zero. If this model is rejected, nonindependence needs to be modeled. The author suggests two perspectives from which nonindependence can be modeled. These two perspectives are explained in the following paragraphs. In the context of  $2 \times 2$  tables, the first perspective is represented, for example, by the odds ratio,

$$\theta = \frac{p_{11} p_{22}}{p_{12} p_{21}} = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}/p_{21}}{p_{12}/p_{22}}$$

Taking this perspective, the log-linear interaction,  $\lambda_{ij}$ , can be described as

$$\lambda_{ij} = G_{ij} - G_{i.} - G_{.j} + G_{..}$$

where

$$G_{ij} = \ln p_{ij}, \\ G_{i.} = \sum_j G_{ij}/J, G_{.j} = \sum_i G_{ij}/I, \text{ and } G_{..} = \sum_i \sum_j G_{ij}/IJ.$$

The second perspective is represented, for example, by the correlation coefficient,

$$\rho = \frac{p_{11} p_{22} - p_{12} p_{21}}{\sqrt{p_{1.} p_{2.} p_{.1} p_{.2}}}$$

Taking the second perspective, one can consider the *relative difference*,

$$\Delta_{ij} = \frac{p_{ij} - p_{i.} p_{.j}}{p_{i.} p_{.j}}$$

From the definition of  $\lambda_{ij}$  follows that

$$\sum_i \lambda_{ij} = 0, j = 1, \dots, J$$

and

$$\sum_j \lambda_{ij} = 0, i = 1, \dots, I.$$

From the definition of the relative difference follows that

$$\sum_i \Delta_{ij} p_{i.} = 0, j = 1, \dots, J,$$

and

$$\sum_j \Delta_{ij} p_{.j} = 0, i = 1, \dots, I.$$

Aggregating over all cells of an  $I \times J$  table yields the following two measures of deviation from independence

$$\lambda = \sqrt{\sum_i \sum_j \frac{\lambda_{ij}^2}{IJ}}$$

and

$$\Delta = \sqrt{\sum_i \sum_j \Delta_{ij}^2 p_{i.} p_{.j}}$$

Clearly,  $\lambda$  and  $\Delta$  reflect different perspectives of deviation from independence. The first perspective is *marginal-free*. The marginal distribution plays no role in the definition and estimation of  $\lambda$ . The second perspective is *marginal-dependent*. The marginals do play a role in the definition and the estimation of  $\Delta$ . Because of these differences, measures that reflect these perspectives can be expected to yield different appraisals of deviation from independence (for examples, see von Eye & Mun, 2003; von Eye, Spiel, & Rovine, 1995).

The relationship of these measures to the odds ratio,  $\theta$ , and Cohen's  $\kappa$  can be demonstrated as follows (for the sake of simplicity, we use the context of  $2 \times 2$  tables, as in the illustration examples in the section "Marginal Dependency of Cohen's  $\kappa$ "). For  $\lambda$ , we find that, in  $2 \times 2$  tables,

$$\lambda = \frac{\ln \theta}{4}.$$

For  $\Delta$ , we find that, in  $2 \times 2$  tables,  $\Delta = |\rho|$ . Considering

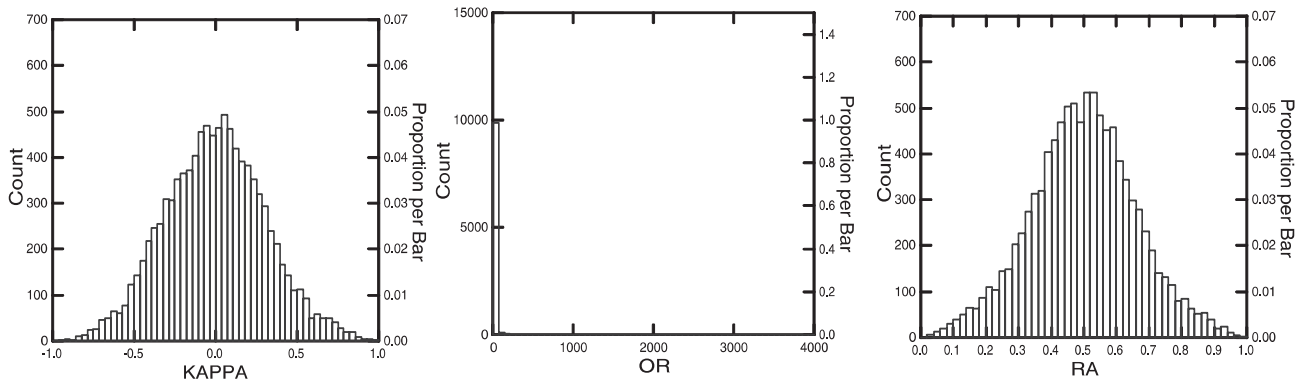


Figure 3. Univariate distributions of  $\kappa$ , the odds ratio  $\theta$  and  $ra$ .

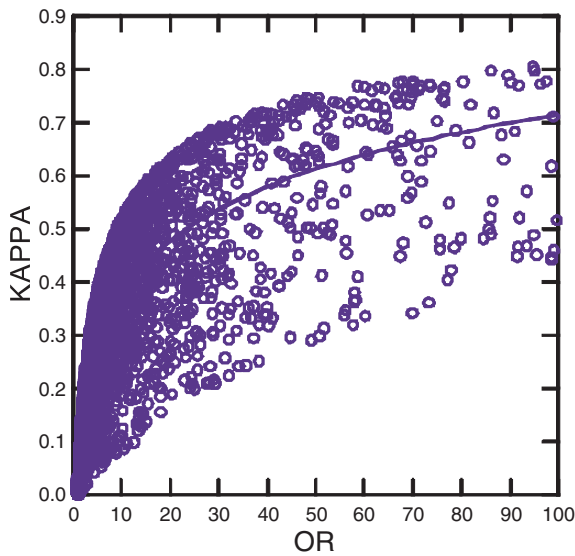


Figure 4. Scatterplot of the odds ratio with  $\kappa$ , for 4906 random  $2 \times 2$  tables, with a logarithmic smoother; for  $\kappa \geq 0$  and  $\theta \leq 100$ .

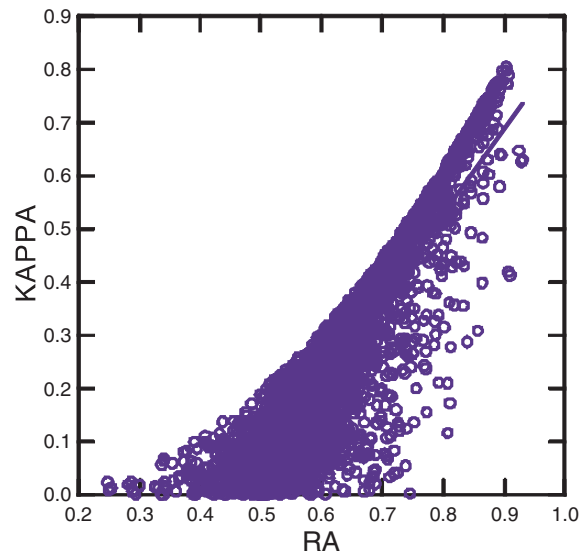


Figure 5. Scatterplot of  $ra$  with  $\kappa$ , for 4906 random  $2 \times 2$  tables, with a linear smoother; for  $\kappa \geq 0$  and  $\theta \leq 100$ .

that  $\theta$  is a simple transformation of  $\lambda$ , we conclude that  $\theta$  is a marginal-free measure of deviation from independence.

The comparison of  $\Delta_{ij}$  with  $\kappa$  shows that, for  $i = j$ , the numerators of the two measures are identical. We, thus, conclude that  $\kappa$  is a marginal-dependent measure of deviation from independence in the diagonal cells. The implication of these comparisons is that, because  $\theta$  is marginal-free and  $\kappa$  is marginal-dependent in Goodman's (1991) sense, the two measures can be expected to yield different appraisals of the frequency distribution in an agreement table.

There are special cases in which marginal-free and marginal-dependent measures yield comparable appraisals of agreement tables. One such case is when  $p_i = 1/I$ , and  $p_j = 1/J$ . In addition, under these conditions, Cohen's  $\kappa$  and Brennan and Prediger's (1981)  $\kappa$  will give comparable appraisals of rater agreement.

### Simulation Results

The examples in the literature and the ones in the section "Explaining the Marginal Dependency of  $\kappa$ " demonstrate specific cases. However, they cannot be used to show whether the selected characteristics are general in nature or restricted to the specific case. Therefore, we performed a simulation study. In this study, 10,000  $2 \times 2$  tables were created using the uniform random number generator available in MATLAB. The function `unidrnd()` generated numbers were restricted to range from 1 through 100, for each cell. The maximum sample size thus ranged from 4 through 400. For each of the thus created cells, the three measures  $\kappa$ ,  $\theta$ , and  $ra$  were calculated. Figure 3 displays the histograms for the three measures.

Figure 3 shows that  $\kappa$  and  $ra$  are approximately normally distributed, as expected. The mean of  $\kappa$  is zero, and the mean of  $ra$  is 0.5. By far the majority of the values of the

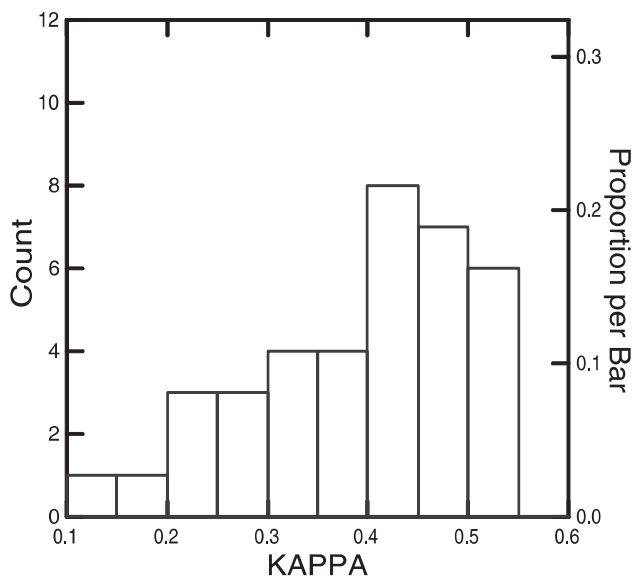


Figure 6. Histogram of  $\kappa$ , for odds ratios  $9.8 \geq \theta \geq 10.2$ .

odds ratio  $\theta$  are between 0 and 250. However, there are a few very large values. The maximum value found in the simulated data was 3,645. The theoretically largest value would have been 10,000.

For the following discussion, we restricted the values of  $\kappa$  and  $\theta$  to values  $\kappa \geq 0$  and  $\theta < 100$ . The first restriction was set because (1)  $\kappa$  is mostly of interest when raters agree to a larger degree than expected, and (2) when  $\kappa$  is negative, it is no longer monotonic (von Eye & Sørensen, 1991). The restriction for  $\theta$  was set because the very extreme values distort the graphical display (and the number of extreme scores was only 4%). The number of tables that remained was 4906. For these tables, the correlations among the three measures were  $r_{\kappa, \theta} = 0.69$ ,  $r_{\kappa, ra} = 0.91$ , and  $r_{\theta, ra} = 0.61$ . Figure 4 displays the scatterplot of  $\theta$  with  $\kappa$ , with a logarithmic smoother.

Figure 4 shows that the relationship between the odds ratio and  $\kappa$  is, over the entire range of the graph (20 values were omitted; for these, the odds ratio was greater than 100 and  $\kappa$  was at or above 0.8), monotonic but non-linear. In addition, for none of the odds ratios  $\theta > 3$ ,  $\kappa$  was zero. For odds ratios  $\theta > 10$ , the range of  $\kappa$  values was about 0.45, with a decreasing tendency for very large odds ratios.

Figure 5 displays the scatterplot of  $ra$  with  $\kappa$ , with a linear smoother. The figure shows that the relationship between  $ra$  and  $\kappa$  is almost linear. For values of  $\kappa = 0$ , a wide range of values of  $ra$  can be observed. Indeed, when  $ra$  indicates disagreement ( $ra < 0.5$ ),  $\kappa$  can still indicate agreement beyond expectation of up to  $\kappa = 0.2$ . As  $ra$  increases,  $\kappa$  also increases, and the variance about the regression line becomes smaller.

Figure 6 extends the example in the section "Marginal Dependency of  $\kappa$  for Constant Association Values." It shows the range of  $\kappa$  values for an odds ratio of  $9.8 \leq \theta \leq 10.2$ .

Figure 6 shows that the range of scores that  $\kappa$  assumed in the present simulations includes, for  $\theta$  of about 10,  $\kappa = 0.109$  and  $\kappa = 0.518$ . Appendix B shows how to calculate  $\kappa$  under the present restriction.

## Discussion

It is well-known and has been demonstrated here again that Cohen's  $\kappa$  will not approximate its maximum value of 1 when the marginal distributions are not uniform. Instead,  $\kappa$  will approximate an asymptote that is clearly less than 1. An exception to this rule is the case in which all disagreement cells have probabilities of zero. In addition,  $\kappa$  has been shown to respond to variations in marginal distributions when other measures – for example, the odds ratio and the measure of raw agreement – remain constant. To summarize this behavior of  $\kappa$ , it is called *marginal dependent*.

To explain this behavior of  $\kappa$ , two arguments were used in this article. The first is that the chance model used for  $\kappa$  is the log-linear main effect model. This model is supposed to take main effects into account. These effects come in the form on nonuniform marginal distributions. Differences in marginal distributions can, therefore, be expected to result in different appraisals of data characteristics, even when variable interactions remain unchanged. The second argument was that measures of deviation from independence can be classified in those that are marginal-free versus those that are marginal-dependent (Goodman, 1991; see Berger & Zhang, 2004). The odds ratio is prototypical of marginal-free measures. The  $\Delta_{ij}$  measure is prototypical of marginal-dependent measures.  $\kappa$  and  $\Delta_{ij}$  (and  $\chi^2$ ) use the same chance model for the estimation of expected cell frequencies. Therefore,

1.  $\kappa$  is marginal-dependent, and
2.  $\kappa$  can be expected to result in different descriptions of data than  $\theta$  or  $ra$ .

Considering these characteristics of  $\kappa$ , we now ask what the basis of the controversy is that surrounds the utility of  $\kappa$  (and weighted  $\kappa$ ; Cohen, 1968). We see one major reason for this controversy. From the context of rules of thumb that have been proposed to categorize the magnitude of  $\kappa$ , we conclude that many researchers tend to interpret  $\kappa$  as a measure of *degree of interrater agreement*. Clearly,  $\kappa$  is not such a measure (unless the marginal distribution is uniform, or  $p_{ij} = 0$  for  $i \neq j$ ). Instead,  $\kappa$  is a measure of agreement beyond a particular chance model. Specifically,  $\kappa$  is a measure of the degree to which the agreement cells contain more cases than expected under the main effect model of rater independence. If  $\kappa$  is high, one can conclude that more cases were found in the agreement cells than expected under this model. This does usually also imply that there are large numbers of cases in the agreement cells. Conversely, when  $\kappa$  is low,

this does not imply that the number of judgments in the agreement cells is low. In the latter case, the observed number of agreements is not much different than the expected number, no matter how strong agreement is.

In short:  $\kappa$  performs as designed. The controversy results from misguided attempts to interpret the coefficient.

As a matter of course, one can challenge other characteristics of  $\kappa$ . For example, one can ask whether the selection of the main effect model as a chance model is appropriate, or what the dependence of  $\kappa$  is on a latent binary variable that represents the true status of a rating object. This and other discussions are interesting and important (see Vach, 2004; von Eye, 2008). However, they operate outside the scope of the present discussion of marginal dependence.

Finally, we ask whether recommendations can be given to researchers who are interested in expressing the degree of rater agreement in one or a few measures. We propose two lines of action:

1. When the marginal distributions are uniform,  $\kappa$  can be used as a measure of both, weighted deviation from a chance model in the agreement cells, and degree of agreement.
2. When the marginal distributions are not uniform,  $\kappa$  can only be used as a measure of deviation from a chance model, and at least one second measure should be used to describe the degree of agreement. Prime candidates for such measures include the coefficient of raw agreement (which equals the quantity  $\theta_1$ ) and Brennan and Prediger's  $\kappa_r$ , which is not marginal dependent in Cohen's sense but can be interpreted as a measure of deviation from a null model – that is, as a measure that describes the degree of agreement.

## Acknowledgments

This paper is based on a presentation given by the first author as part of an invited Workshop at the 10th European Congress of Psychology held in Prague, Czech Republic, July 2007.

## References

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.
- Berger, V.W., & Zhang, J. (2005). Marginal independence. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1126–1128). Chichester, UK: Wiley.
- Bortz, J., & Lienert, G.A. (1998). *Kurzgefaßte Statistik für die klinische Forschung* [A short introduction to statistics for clinicians]. Berlin, Germany: Springer-Verlag.
- 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.
- Cook, R.J. (1998).  $\kappa$  and its dependence on marginal rates. In T.P. Armitage & T. Colton (Eds.), *The encyclopedia of biostatistics* (pp. 2166–2168). New York: Wiley.
- Cook, R.J., & Farewell, V.T. (1995). Conditional inference for subject-specific and marginal agreement: Two families of agreement measures. *The Canadian Journal of Statistics*, *23*, 333–344.
- Fleiss, J.L. (1975). Measuring agreement between two judges in the presence or absence of a trait. *Biometrics*, *31*, 651–659.
- Fleiss, J.L., Levin, B., & Paik, M.C. (2003). *Statistical models for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley.
- Froman, T., & Llabre, J.H. (1985). The equivalence of  $\kappa$  and Del. *Perceptual and Motor Skills*, *60*, 651–659.
- Goodman, L.A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association*, *86*, 1085–1111.
- Goodman, L.A., & Kruskal, W.H. (1954). Measures of association for cross-classifications. *Journal of the American Statistical Association*, *49*, 732–764.
- Guggenmoos-Holzmann, I. (1995). Modeling covariate effects in observer agreement studies: The case of nominal scale agreement (letter to the editor). *Statistics in Medicine*, *14*, 2285–2286.
- Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.
- Mun, E.Y. (2005). Rater agreement –  $\kappa$ . In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1712–1714). Chichester, UK: Wiley.
- Vach, W. (2005). The dependence of Cohen's  $\kappa$  on the prevalence does not matter. *Journal of Clinical Epidemiology*, *58*, 655–661.
- von Eye, A. (2008). *What do we know about Cohen's  $\kappa$ ? A review and discussion*. In M. Stemmler, E. Lautsch, & B. Martinke (Eds.), *Configural frequency analysis (CFA) and other non-parametric methods* (pp. 29–39). Lengerich: Papst Science Publishers.
- von Eye, A., & Mun, E.Y. (2003). Characteristics of measures for  $2 \times 2$  tables. *Understanding Statistics*, *2*, 243–266.
- von Eye, A., & Mun, E.Y. (2005). *Modeling rater agreement – manifest variable approaches*. Mahwah, NJ: Erlbaum.
- von Eye, A., Schauerhuber, M., & Mair, P. (2007). Significance tests for the measure of raw agreement. *InterStat*. Retrieved January 2007, from <http://interstat.statjournals.net/YEAR/2007/abstracts/0701001.php> [Also reproduced in <http://epub.wu-wien.ac.at/dyn/virlib/metasearch/advancedquery?V1=von+Eye&errors=w&back=%2F>]
- von Eye, A., & Sörensen, S. (1991). Models of chance when measuring interrater agreement with  $\kappa$ . *Biometrical Journal*, *33*, 781–787.
- von Eye, A., Spiel, C., & Rovine, M.J. (1995). Concepts of non-independence in configural frequency analysis. *Journal of Mathematical Sociology*, *20*, 41–54.
- von Eye, A., & von Eye, M. (2005). Can one use Cohen's  $\kappa$  to examine disagreement? *Methodology*, *1*, 129–142.
- Wickens, T. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Erlbaum.

## About the authors

Alexander von Eye is professor at Michigan State University. His research focuses on the development and application of statistical methods, including methods for the analysis of categorical data, longitudinal data, classification, computational statistics, and structural equation modeling.

Maxine von Eye is currently working on her Ph.D. in mathematics at the University of Cambridge, UK. Her research focuses on mathematical modeling, oceanography, time series, physical experimentation, and computer simulation.

## Alexander von Eye

Michigan State University  
316 Psychology Building  
East Lansing, MI 48864-1116  
USA  
E-mail voneye@msu.edu

## Appendix A

The probability of cells of a  $2 \times 2$  table for given, symmetric marginals

Let

- 1)  $p_{1.} = \beta$ ,
- 2)  $p_{.1} = \beta$ ,
- 3)  $\sum p_{ii} = 1$ , and
- 4) the odds ratio  $(p_{11} p_{22}) / (p_{12} p_{21}) = \theta$ .

Subtracting (1) from (2) results in

5)  $p_{12} = p_{21}$ . Solving (1) for  $p_{11}$  and substituting in (5) yields

6)  $p_{11} = 1 - p_{12} - p_{22}$ . Solving (2) for  $p_{11}$  results in

7)  $p_{11} = \beta - p_{12}$ . Equating (6) and (7) to solve for  $p_{22}$  yields

8)  $p_{22} = (1 - \beta) - p_{12}$ . Substituting (7) and (8) into (4) and rearranging to get quadratic equation for  $p_{12}$  results in  $(1 - \theta) p_{12}^2 - p_{12} + \beta - \beta^2 = 0$ . This quadratic equation has the solution

$$p_{12} = 1 \pm \frac{\sqrt{1 - 4(\beta - \beta^2)(1 - \theta)}}{2(1 - \theta)},$$

of which the values  $0 \leq p_{12} \leq 1$  are selected. The solutions for  $p_{12}$  are real if

$$1 - 4(\beta - \beta^2)(1 - \theta) \geq 0 \Leftrightarrow \theta \geq 1 - 1/(4(\beta - \beta^2)).$$

## Appendix B

Calculation of  $\kappa$  for given, symmetric marginals

$\kappa$  is defined by

$$\kappa = \frac{\sum_i p_{ii} - \sum_i p_{i.} p_{.i}}{1 - \sum_i p_{i.} p_{.i}}.$$

For  $p_{1.} = \beta$  and  $p_{.1} = \beta$ , and  $p_{12}$  (defined in Appendix A), one obtains

$$\kappa = \frac{-p_{12} + \beta - \beta^2}{\beta - \beta^2}.$$