J Clin Epidemiol Vol. 46, No. 5, pp. 423–429, 1993
Printed in Great Britain.

0895-4356/93 $6.00 + 0.00

BIAS, PREVALENCE AND KAPPA

TED BYRT,* JANET BISHOP and JOHN B. CARLIN
Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital Research Foundation and Melbourne University Department of Paediatrics, Melbourne, Australia

(Received in revised form 4 January 1993)

Abstract—Since the introduction of Cohen's kappa as a chance-adjusted measure of agreement between two observers, several "paradoxes" in its interpretation have been pointed out. The difficulties occur because kappa not only measures agreement but is also affected in complex ways by the presence of bias between observers and by the distributions of data across the categories that are used ("prevalence"). In this paper, new indices that provide independent measures of bias and prevalence, as well as of observed agreement, are defined and a simple formula is derived that expresses kappa in terms of these three indices. When comparisons are made between agreement studies it can be misleading to report kappa values alone, and it is recommended that researchers also include quantitative indicators of bias and prevalence.

Kappa Agreement Bias Prevalence

INTRODUCTION

Suppose two raters classify N subjects as belonging to one of two categories such as Yes/No (for example, two radiologists classifying N X-rays as tumor present/tumor absent). The results can be arranged in a 2×2 table as follows:

		Observer A		
		Yes	No	Total
Observer B	Yes	a	b	g_1
	No	c	d	g_2
	Total	f_1	f_2	N

The proportion of observed agreement is $p_o = (a + d)/N$. Some of this agreement can be expected by chance alone, in the sense that two observers classifying subjects quite independently (each with a given proportion of "Yes"s) will by chance agree on a predictable proportion of cases. There have been many proposals for interrater agreement coefficients which aim to measure chance-corrected agreement, the most influential of these being Cohen's kappa [1].

Zwick [2] examines three coefficients of agreement each of which can be expressed in the form

$$\frac{p_o - p_e}{1 - p_e}$$

where p_e denotes the agreement expected by chance. The three different coefficients can be characterized by their definition of p_e. In Bennett *et al.*'s [3] S coefficient p_e is taken to be 0.5, in which case S is merely a linear transformation of p_o and so incorporates no adjustment for chance. Zwick points out that this coefficient is equivalent to those proposed by a number of others, including Brennan and Prediger [4] and Maxwell [5].

A second agreement coefficient of this form is Scott's [6] π which takes

$$p_e = \frac{(f_1 + g_1)^2 + (f_2 + g_2)^2}{4N^2}.$$

Cohen's kappa (K) is also of this form, taking $p_e = (f_1 g_1 + f_2 g_2)/N^2$. Fleiss [7] comments on the desirable properties of K as a measure of agreement:

> "If there is complete agreement, K = +1. If observed agreement is greater than or equal to

*All correspondence should be addressed to: E. Byrt, Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital, Parkville, Victoria, Australia 3052.

chance agreement, $K \geqslant 0$, and if observed agreement is less than or equal to chance agreement, $K \leqslant 0$. The minimum value of K depends on the marginal proportions."

Although Cohen's kappa has been widely used as a simple single index of agreement, several authors have identified difficulties or paradoxes associated with its interpretation [2, 4, 8–13]. In this paper we review the so-called paradoxes and propose a new framework for examining them, using the key concepts of prevalence (proportion of "Yes" responses) and bias (difference between observers' proportions of "Yes" responses). The discussion focuses entirely on the simplest case, of two raters with a dichotomous classification, but extension to more general situations can be made.

TWO PARADOXES ASSOCIATED WITH KAPPA

Paradox 1 is described by Feinstein and Cicchetti [8]:

"If p_e is large, the [chance] correction process can convert a relatively high value of p_o into a relatively low value of K ... Thus, with different values of p_e, the K for identical values of p_o can be more than twofold higher in one instance than the other."

They provide the example of Tables 1 and 2, both of which have $p_o = 0.85$, while K = 0.70 for the first table and 0.32 for the second.

Paradox 2 is also described by Feinstein and Cicchetti: "Unbalanced marginal totals produce higher values of K than more balanced totals". They provide two examples of this paradox, one of which is reproduced in Tables 3 and 4.

In Table 3 K = 0.13 and in Table 4 K = 0.26, so Table 4 produces a higher kappa despite the fact that its marginal totals show poorer agreement. Note again that the observed agreement is the same for both tables (0.6). Zwick [2] refers to a similar case and observes that two judges who produce identical marginal distributions are penalized when compared with judges who produce different marginals, and "this appears to be an undesirable property".

These paradoxes occur because kappa is affected both by any bias between the raters and also by the overall prevalence (the relative probability of the responses "Yes" and "No").

BIAS, PREVALENCE AND ADJUSTED KAPPAS

Bias

If Observers A and B differ in their assessment of the frequency of occurrence of a condition in a study group, we say that there is a bias between the observers. When this occurs the marginal distributions for the raters are unequal. We define the Bias Index (BI) to be equal to the difference in proportions of "Yes" for the two raters and estimate it by $(a + b)/N - (a + c)/N$, i.e. by $(b - c)/N$. The absolute value of BI has a minimum of 0 when $b = c$, and a maximum of 1 when either $b = N$ or $c = N$. The Bias Index is equal to zero if and only if the marginal proportions are equal. In this case $f_1 = g_1$ and $f_2 = g_2$.

The effect of bias can be seen when comparing Tables 5 and 6.

The value of kappa for Table 5, where both observers classify 60% of cases "Yes", is 0.17 and for Table 6, where there are different marginal proportions (45% and 75%) is 0.24. In the first case BI = 0 while in the second BI = 0.3.

Now we define a version of kappa that adjusts for bias. Bias-adjusted kappa (BAK) is the value of kappa that results if b and c in our original table are both replaced by their average, $m = (b + c)/2$. BAK is in fact Scott's [6] π, although we derive it from a different point of view. For both Table 5 and Table 6, BAK = 0.17; the value of kappa is higher for Table 6 entirely because of the bias between the observers (as bias increases, p_e declines and K increases: see Appendix A).

Table 1 kappa=0.70

		Observer A		
		Yes	No	Total
Observer B	Yes	40	9	49
	No	6	45	51
	Total	46	54	100

p0 = 0.85 (tables 1 e 2)

Paradoxo 1: diferença nas prevalências gera kappas menores

Table 2 kappa=0.32

		Observer A		
		Yes	No	Total
Observer B	Yes	80	10	90
	No	5	5	10
	Total	85	15	100

PI grande (prevalência das categorias muito diferente)

Table 3 kappa=0.13

		Observer A		
		Yes	No	Total
Observer B	Yes	45	15	60
	No	25	15	40
	Total	70	30	100

p0 = 0.60 (tables 3 e 4)

Paradoxo 2: Desbalanceamento das marginais gera kappas maiores

Table 4 kappa=0.26

		Observer A		
		Yes	No	Total
Observer B	Yes	25	35	60
	No	5	35	40
	Total	30	70	100

BI grande (% de classificação nas categorias muito diferente entre observadores)

$$p_e = \frac{(f_1 + g_1)^2 + (f_2 + g_2)^2}{4N^2}.$$

Table 5 kappa=0.17

		Observer A		
		Yes	No	Total
Observer B	Yes	40	20	60
	No	20	20	40
	Total	60	40	100

p0=0.60 (tables 5 e 6)

Paradoxo 2: Desbalanceamento das marginais gera kappas maiores

Table 6 kappa=0.24

		Observer A		
		Yes	No	Total
	Yes	40	35	75
Observer B	No	5	20	25
	Total	45	55	100

Table 7

		Observer A		
		Yes	No	Total
Observer B	Yes	40	10	50
	No	10	40	50
	Total	50	50	100

p0=0.60 (tables 7 e 8)

Paradoxo 1: diferença nas prevalências gera kappas menores

Table 8

		Observer A		
		Yes	No	Total
Observer B	Yes	70	10	80
	No	10	10	20
	Total	80	20	100

Prevalence

The value of kappa is also affected by the relative probabilities of the "Yes" and "No" categories. If we wished to estimate the probability of "Yes" for the whole population, the best estimate, from our sample, would be the mean of $(a + b)/N$ and $(a + c)/N$. Similarly, the best estimate of the probability of "No" is the mean of $(c + d)/N$ and $(b + d)/N$. We call the difference between the probability of "Yes" and the probability of "No" the Prevalence Index (PI). It is estimated by $(a - d)/N$, which takes values from -1 (when $a = 0$ and $d = N$) to $+1$ (when $a = N$ and $d = 0$) and is equal to 0 when "Yes" and "No" are equally probable, i.e. when the average prevalence of "Yes" is 50%.

For example, consider the cases in Tables 7 and 8.

In both cases there are 80 agreements and 20 disagreements between the raters, but in the first case K = 0.6 and PI = 0, while in the second K = 0.375 and PI = 0.6. The large difference between kappa values is due entirely to the prevalence effect—the larger the value of PI, the larger is p_e, and the smaller is K (Appendix A).

Prevalence - adjusted bias - adjusted kappa (PABAK)

It is illuminating to define an index of agreement between two observers (denoted PABAK) that adjusts kappa for differences in prevalence of the conditions "Yes" and "No", and for bias between observers, by obtaining the value of kappa that results when we not only replace b and c by their average, $m = (b + c)/2$, but also replace a and d by their average, $n = (a + d)/2$. In other words we compute PABAK as the value of kappa from the following table:

		Observer A	
		Yes	No
Observer B	Yes	n	m
	No	m	n

It can easily be seen that in this table $p_e = 0.5$, so we obtain

$$\text{PABAK} = \frac{(2n/N) - 0.5}{1 - 0.5} = 2p_o - 1.$$

It is revealing that a natural process of adjusting for bias and prevalence leads to an index linearly related to the observed agreement, p_o. In fact, PABAK is the same as Bennett's S coefficient, but as with BAK, our derivation is different. PABAK rescales p_o so it takes values from -1 (when $n = 0$) to $+1$ (when $m = 0$) and is zero when observed agreement is equal to 50%.

The combined effects of bias and prevalence

It is shown in Appendix B that kappa is related to PABAK by the following formula:

$$\text{K} = \frac{\text{PABAK} - \text{PI}^2 + \text{BI}^2}{1 - \text{PI}^2 + \text{BI}^2}. \qquad (1)$$

It can be seen that, unless PABAK = 1, the larger the absolute value of BI, the larger is K (for PI constant), and the larger the absolute value of PI, the smaller is K (for BI constant). If both bias and prevalence effects are present, then the result may be that K is larger or smaller than PABAK, depending on the relative size of BI and PI. The dependence of kappa on PI and BI, for given PABAKs, can be seen clearly in Fig. 1.

Note that BI and PI may vary quite independently of each other, the former reflecting difference between the cells of disagreement, b and c, and the latter difference between the cells of agreement, a and d. The value of PABAK or p_o determines maximum possible values of BI and PI, and for any p_o these maximum values sum to one.

The potential effect of bias is much greater for small values of K than for large values. When PABAK is 0.8 ($p_o = 0.9$) the largest possible BI

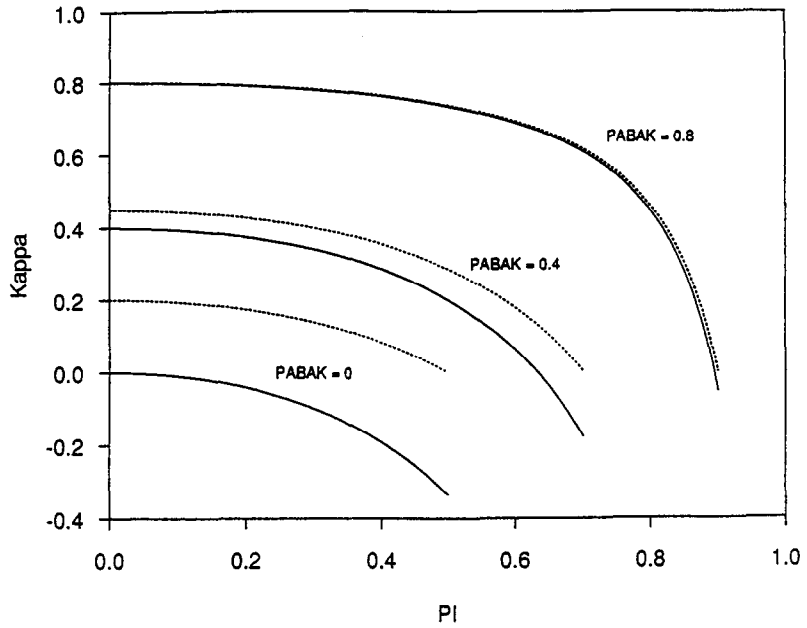

Fig. 1. Relationship of kappa to PI for three values of PABAK, as marked, and in each case for a bias of zero (solid line) and for the maximum possible bias (dotted line).

is 0.1, so the bias effect is necessarily small. However, when PABAK is close to zero BI can be as high as 0.5, producing a difference between PABAK and K of as much as 0.2. On the other hand, the potential prevalence effect is much greater for large values of PABAK or p_o than for small values. For example, in an extreme case, such as Table 9, there is an observed agreement of 90% and PABAK of 0.8, but K is negative. Figure 2 illustrates the potential joint effects of bias and prevalence on kappa.

When there is both a prevalence effect and a bias effect it can be difficult to distinguish between the effects, and it is this difficulty that produces the so-called "paradoxes". For example, returning to Tables 1 and 2 (Paradox 1), there is only a small bias effect in each table (BI = 0.03, 0.05, respectively) but there is a large differential prevalence effect (PI = 0.05, 0.75), which results in Table 2 producing a much smaller K than Table 1. In Tables 3 and 4 (Paradox 2) there are moderate prevalence effects (PI = 0.3, 0.1, respectively) that are overwhelmed by a stronger bias effect (BI = −0.1, 0.3), resulting in Table 4 having a larger K than Table 3. Table 10 provides a summary of the various indices for Tables 1–4.

Table 9

		Observer A	
		Yes	No
Observer B	Yes	90	5
	No	5	0

DISCUSSION

We have shown that for a 2 × 2 table of agreement kappa can be simply expressed in terms of three easily interpreted indices, PABAK (or alternatively p_o), which is a measure of observed agreement, BI, an index of the bias between observers, and PI, an index of the differences between the overall proportion of "Yes" and "No" assessments. This reexpression of kappa enables a clear explanation of the conceptually distinct and independent components that arise in the investigation of agreement. For example, Paradoxes 1 and 2 above can be completely explained in terms of the prevalence effect (increasing PI decreases kappa) combined with the bias effect (increasing BI increases kappa).

Zwick [2] has gone part-way to identifying the need to distinguish between agreement, prevalence and bias by describing three cases in each of which p_o is 0.60, but kappa has values of 0.467, 0.444 and 0.474, respectively. In the first case the marginals are uniform ($f_1 = g_1 = f_2 = g_2$), in the second case the

Table 10

	p_o	BI	PI	K	BAK	PABAK
Table 1	0.85	0.03	−0.05	0.70	0.70	0.70
Table 2	0.85	0.05	0.75	0.32	0.31	0.70
Table 3	0.60	−0.10	0.30	0.13	0.12	0.20
Table 4	0.60	0.30	0.10	0.26	0.19	0.20

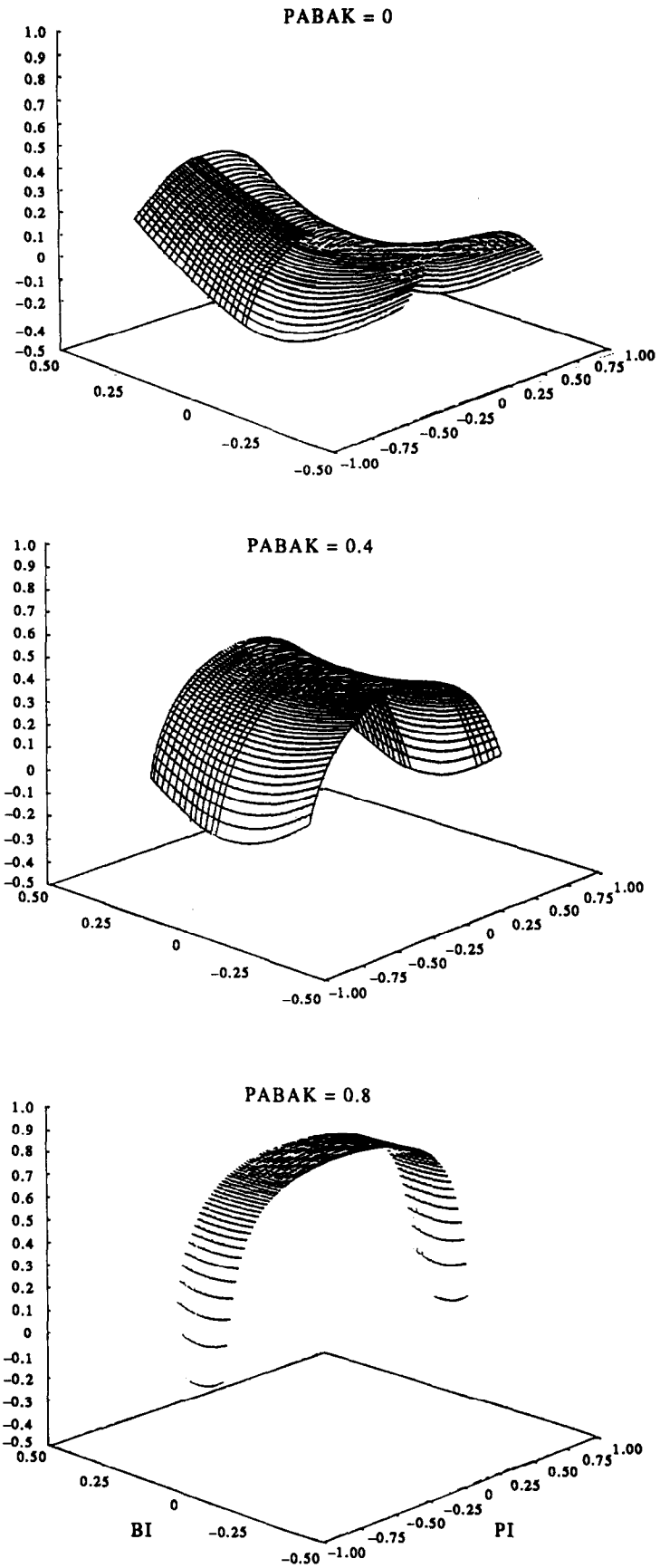

Fig. 2. Relationship of kappa to BI and PI, simultaneously, for three values of PABAK.

marginals are equal but not uniform ($f_1 = g_1$ and $f_2 = g_2$) and in the third case the marginals are unequal. The three examples correspond to situations in which (a) there is neither a prevalence effect nor bias, (b) there is a prevalence effect but no bias, so the value of kappa is reduced, and (c) there is a prevalence effect and bias and, since the bias effect is the stronger, the value of kappa is raised.

The bias effect can be large only when there is poor agreement and low kappa (Fig. 1). If PABAK = 0 (observed agreement = 50%) and there is no prevalence effect (prevalence = 50%, PI = 0), kappa ranges from 0, in the case of no bias, to 0.2, which represents "slight agreement" on the scale of Landis and Koch [14], in the case of maximal bias. Despite the generally small magnitude of the effect, several authors have expressed reservations about the counter-intuitive effect of bias on kappa [2, 4].

In addition to the effect on kappa, it will often be of interest to know about the existence of bias for its own sake. First, when assessing agreement between something and an accepted criterion, we should be particularly interested to know about the presence and magnitude of any bias. Second, if we believe we have interchangeable raters and have no reason to expect that bias will be present, then the unexpected occurrence of bias alerts us to something that may be important and should be investigated further.

A number of authors have commented on the effect of prevalence in the interpretation of kappa statistics [8–13]. Kraemer [13] has specifically observed that it is possible that something that would be regarded as poor reliability in terms of the standards suggested by Landis and Koch [14] can result from low prevalence and hence reflect the nature of the population rather than the observation procedure. Although Whitehurst [10] regards such a situation as objectionable, Zwick [2] considers that it is not unreasonable that the value of an agreement coefficient should be smaller in a case where the marginals are equal but not uniform, than in a case where there are uniform marginals. The dependence of kappa on prevalence is not necessarily bad, since there is more likely to be chance agreement when the prevalence is high or low than when it is 50%. Researchers should be aware, however, that the prevalence effect can be very large, for example as in Table 10, and, as with bias effects, if kappa alone is reported it is misleading to compare kappas for raters between situations where the average prevalence is considerably different.

Table 11

35	15	33	34	86	5	50	30
15	35	0	33	5	4	0	20

A reported value of kappa = 0.4 can mean many different things. For example, it can mean an observed agreement of 0.7 with BI = 0 and PI = 0, an observed agreement of 0.66 with BI = 0.34 and PI = 0, an observed agreement of 0.9 with BI = 0 and PI = 0.82, or an observed agreement of 0.7 with BI = 0.30 and PI = 0.30. This is illustrated in Table 11, where each of the four tables has K = 0.4.

Cicchetti and Feinstein [9] have suggested that Paradoxes 1 and 2 may be resolved by calculating the observed proportions of positive and negative agreement, $p_{pos} = 2a/(N + a - d)$ and $p_{neg} = 2d/(N - a + d)$. These indices are both closely related to our prevalence index: $p_{pos} = (p_o + PI)/(1 + PI)$ and $p_{neg} = (p_o - PI)/(1 - PI)$, but they do not reflect the effect of bias. For example, in Tables 5 and 6, where the bias indices (BI) are 0 and 0.3 respectively, $p_{pos} = 2/3$ and $p_{neg} = 1/2$ for both tables. We feel that it is important to consider the effects of bias and prevalence independently.

We endorse Cicchetti and Feinstein's observation that no single omnibus index of agreement can be satisfactory for all purposes. As a strategy for interpreting 2 × 2 agreement tables and reporting results we recommend the following:

(1) The presence of bias should be assessed using the bias index BI, interpreted both in substantive terms (is this amount of bias important in this particular context?) and in terms of its effect on kappa. If substantial bias is present, this should be investigated further to discover its cause, and it may be inappropriate or unnecessary to quote an index of agreement.

(2) The effect of prevalence should be assessed using the prevalence index, PI, in conjunction with Fig. 1. If bias effects are relatively small, which is often the case, the interpretation of kappa should be qualified explicitly by some discussion of the prevalence effect. For practical interpretation of prevalence effects it may be helpful to use Cicchetti and Feinstein's indices of positive and negative agreement, p_{pos} and p_{neg}.

In conclusion, like many others, we have observed that the reporting of a single coefficient of agreement makes interpretation and comparison difficult, and have shown how kappa may be decomposed into components reflecting observed agreement, bias and prevalence. When comparisons are made between agreement studies it can be misleading to report kappa values alone, and it is recommended that researchers should also discuss the effects of bias and prevalence.

REFERENCES

1. Cohen JA. A coefficient of agreement for nominal scales. **Educ Psychol Meas** 1960; 20: 37–46.
2. Zwick R. Another look at interrater agreement. **Psychol Bull** 1988; 103: 374–378.
3. Bennett EM, Alpert R, Goldstein AC. Communications through limited response questioning. **Public Opinion Q** 1954; 18: 303–308.
4. Brennan RL, Prediger D. Coefficient kappa: Some uses, misuses, and alternatives. **Educ Psychol Meas** 1981; 41: 687–699.
5. Maxwell AE. Coefficients of agreement between observers and their interpretation. **Br J Psychiatry** 1977; 116: 651–655.
6. Scott WA. Reliability of content analysis: The case of nominal scale coding. **Public Opinion Q** 1955; 19: 321–325.
7. Fleiss JL. **Statistical Methods for Rates and Proportions**, 2nd edn. New York: Wiley; 1981: 217.
8. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The Problems of Two Paradoxes. **J Clin Epidemiol** 1990; 43: 543–548.
9. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. **J Clin Epidemiol** 1990; 43: 551–558.
10. Whitehurst GJ. Interrater agreement for journal manuscript reviews. **Am Psychol** 1984; 39: 22–28.
11. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. **Am J Epidemiol** 1987; 126: 161–169.
12. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. **J Clin Epidemiol** 1988; 41: 949–958.
13. Kraemer HC. Ramifications of a population model for kappa as a coefficient of reliability. **Psychometrika** 1979; 44: 461–472.
14. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. **Biometrics** 1977; 33: 159–174.

APPENDIX A

For a fixed observed agreement, p_o, the larger the expected agreement, the smaller is the value of kappa.

Proof:

Kappa is defined by

$$K = \frac{p_o - p_e}{1 - p_e} \text{ which equals } 1 - \frac{1 - p_o}{1 - p_e}.$$

As p_e increases, $1 - p_e$ decreases,
$1/(1 - p_e)$ increases,
and K decreases.

APPENDIX B

For any 2 × 2 table,

$$a = N(p_o + \text{PI})/2$$
$$b = N(1 - p_o + \text{BI})/2$$
$$c = N(1 - p_o - \text{BI})/2$$
$$d = N(p_o - \text{PI})/2$$

and it follows that

$$K = \frac{\text{PABAK} - \text{PI}^2 + \text{BI}^2}{1 - \text{PI}^2 + \text{BI}^2}, \text{ where PABAK} = 2p_o - 1.$$