Psychological Bulletin
1987, Vol. 102, No. 2, 315-321

0033-2909/87/$00.75

Standard Error of the Kappa Statistic

James A. Hanley
Department of Epidemiology and Biostatistics
McGill University, Montreal, Quebec, Canada

I show that the large-sample standard error of kappa is largely determined by just two parameters, kappa itself and the proportion of agreement expected by chance. I provide nomograms relating the standard error to these two parameters. These nomograms can be used to anticipate the degree of precision provided by a given sample size or to determine the sample size required for a prespecified level of precision. Also, they are sufficiently accurate over the range of interest that they can be used, instead of the usual lengthier formula, to obtain the standard error of the observed kappa.

The kappa coefficient (κ) is a measure of association used to describe the degree of interrater agreement when using a nominal scale. It plays a role for nominal measures analogous to that played by the intraclass coefficient for interval measures and can be used as one criterion of the validity of a nominal scale (Kraemer, 1983).

In its original and simplest form (Cohen, 1960), κ can be calculated when two fixed observers (or two methods) independently classify the same N (randomly chosen) subjects into k mutually exclusive categories. As is shown in Table 1, these assessments produce marginal proportions $\{p_{i.}\}$ and $\{p_{.j}\}$ (the proportions of subjects classified into each category by each of the two observers separately), along with a $k \times k$ table, giving the proportions p_{ij} of the number of subjects who are classified into category i by the first observer and into category j by the second. The proportion of subjects on whom the observers agree, often called the observed agreement, or p_o, is then $\Sigma\, p_{ii}$. It can be argued that if the observers classified subjects in the proportions $\{p_{i.}\}$ and $\{p_{.j}\}$, respectively, but had no good reason to agree on any particular subject, they would agree, by chance alone, on a proportion $p_e = \Sigma\, p_{i.}p_{.i}$ of subjects, leaving only the remaining fraction $(1 - p_e)$ on which agreement would be more than just by chance. Exactly how much of this potential excess is actually attained is measured by the fraction $\kappa = (p_o - p_e)/(1 - p_e)$, thus making κ a "chance-corrected" measure of agreement, ranging in value from 1 (complete agreement) to 0 (no agreement beyond chance) to a lower limit between 0 and -1 (less than chance agreement).

To some, the kappa coefficient is unnecessarily stringent in crediting so much of observed agreement to chance; if certain categories predominate, seemingly good agreement can still result in low values of kappa (Walter, 1984). In spite of this, the κ statistic has become very popular: Cohen's 1960 paper was cited in over 810 publications in the social sciences between 1960 and 1985 (Institute for Scientific Information, 1986). It has been extended to the case of several fixed or random observers and to the case of ordered categories, with partial credit for partial agreement (Kraemer, 1983). Inferential procedures (standard errors, confidence intervals, and tests of significance) have been developed for most large-sample situations (Fleiss, 1981).

The standard error (SE) of κ is a closed form expression that can be evaluated with a calculator once the data have been tabulated (see Equation 1). In contrast to other statistics based on tabulated data (binomial parameters, odds ratios, and so on), however, the expression is cumbersome, and it is not easy to see how its various components affect the result. Thus, it is difficult to assess, before the data have been collected, what the size of $SE(\kappa)$ is likely to be and what sample size (N) is needed for any desired level of statistical precision or power. Investigators can only make up various scenarios, calculate κ and $SE(\kappa)$ for each, and by trial and error arrive at an N that ensures the required precision. This approach is time-consuming, particularly if one is interested in the nonnull case ($\kappa > 0$) and if k is greater than 2.

Therefore, I investigated the behavior of the nonnull $SE(\kappa)$, restricting my attention to unweighted κ, in studies with $k = 2$, 3, or 4 and in which there were two fixed observers. The purposes were to (a) study the sources of variation in the magnitude of $SE(\kappa)$ (I hypothesized that the SE is largely determined by κ and by p_e in much the same way that the SE of a binomial proportion, p, is determined by its expected value and by N); (b) present a nomogram, displaying the influence of κ and p_e on $SE(\kappa)$, which could be used both in planning the size of a sample and in calculating the SE of the κ obtained in the actual sample; and (c) if possible, provide an intuitive approach to the calculation of $SE(\kappa)$ by considering κ as an estimate of a binomial parameter derived from a sample with a reduced N.

Method

As explained above, let $\{p_{ij}\}$ denote the proportions of the number of subjects who are classified into category i by the first observer and into category j by the second; the marginal proportions $\{p_{i.}\}$ and $\{p_{.j}\}$ are the proportions of subjects classified into each category by each of the two observers separately. Thus,

This work was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada and carried out with the technical assistance of Shari Caplan and Carl Brewer.

I thank Mona Baumgarten and Sheila Dubois for their suggestions.

Correspondence concerning this article should be addressed to James A. Hanley, Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montréal, Québec H3A 1A2, Canada.

Table 1
Joint Proportions of Classifications, by Two Observers, of a Sample of Subjects Into k Categories

	Second observer			
First observer	1	j	k	Total
1	p_{11}	p_{1j}	p_{1k}	$p_{1\cdot}$
i	p_{i1}	p_{ij}	p_{ik}	$p_{i\cdot}$
k	p_{k1}	p_{kj}	p_{kk}	$p_{k\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot j}$	$p_{\cdot k}$	1

$$p_{i\cdot} = \sum_{j=1}^{k} p_{ij}, \quad \text{and} \quad p_{\cdot j} = \sum_{i=1}^{k} p_{ij}.$$

Then

$$p_o = \sum_{i=1}^{k} p_{ii},$$

$$p_e = \sum_{i=1}^{k} p_{i\cdot} p_{\cdot j}, \text{ and}$$

$$\kappa = (p_o - p_e)/(1 - p_e).$$

As shown by Fleiss, Cohen, and Everitt (1969), the standard error of κ is estimated by

$$SE(\kappa) = (A + B - C)^{1/2}/[(1 - p_e)N^{1/2}], \qquad (1)$$

where

$$A = \sum_{i=1}^{k} p_{ii}[1 - (p_{i\cdot} + p_{\cdot i})(1 - \kappa)]^2,$$

$$B = (1 - \kappa)^2 \sum_{i \neq j}^{k} \sum p_{ij}(p_{i\cdot} + p_{\cdot j})^2, \quad \text{and}$$

$$C = [\kappa - p_e(1 - \kappa)]^2.$$

Table 2
$\sqrt{N} \cdot SE(\kappa)$ as a Function of κ and p_e When k = 2

	κ						
p_e	.3	.4	.5	.6	.7	.8	.9
.9	1.88	1.94	1.91	1.81	1.63	1.37	0.99
	(2.1)	(1.3)	(0.8)	(0.5)	(0.3)	(0.1)	(0.1)
.8	1.33	1.36	1.33	1.26	1.14	0.96	0.70
	(4.2)	(2.7)	(1.7)	(1.0)	(0.6)	(0.3)	(0.1)
.7	1.08	1.10	1.08	1.02	0.92	0.78	0.57
	(6.3)	(4.1)	(2.7)	(1.6)	(0.9)	(0.4)	(0.1)
.6	0.93	0.94	0.92	0.87	0.79	0.67	0.49
	(8.5)	(5.6)	(3.6)	(2.2)	(1.2)	(0.5)	(0.1)
.5	0.86	0.85	0.83	0.78	0.70	0.60	0.44
	(10.6)	(7.2)	(4.8)	(2.9)	(1.6)	(0.7)	(0.2)

Note. The upper entry in each cell is the average of the $\sqrt{N} \cdot SE$s obtained in 20 different 2 × 2 tables, all yielding the same value of κ and p_e; the lower entry, in parentheses, is the (average) percentage discrepancy between this average $\sqrt{N} \cdot SE$ and the 20 individual values of $\sqrt{N} \cdot SE$. The maximum discrepancy from the tabulated $\sqrt{N} \cdot SE$ was generally twice the average discrepancy.

Table 3
Average Percentage Error in Approximating $\sqrt{N} \cdot SE(\kappa)$ by a Function of Only κ and p_e When k = 3 and 4

	κ			
p_e	0.3	0.5	0.7	0.9
.8				
$k = 3$	10%	7%	5%	2%
$k = 4$	18%	11%	6%	6%
.6				
$k = 3$	10%	6%	4%	6%
$k = 4$	15%	8%	4%	6%
.4				
$k = 3$	6%	3%	3%	2%
$k = 4$	17%	6%	3%	8%

Note. The error is computed by comparing the average discrepancy of $\sqrt{N} \cdot SE(\kappa)$ from the average of the $\sqrt{N} \cdot SE(\kappa)$s within a 0.05 × 0.05 cell surrounding the indicated values of κ and p_e.

The objective was to simplify Equation 1 by approximating it in terms of κ, p_e, and as few additional parameters as possible. Two different approaches, one for $k = 2$ and another for $k = 3$ and 4, were necessary.

$k = 2$

When $k = 2$, Equation 1 contains six variables: the four p_{ij}s, κ, and p_e. (This does not count $\sqrt{N}$, which appears, as expected, in the denominator.) Both κ and p_e, however, are derived from the ps, and only three of the latter can vary independently, because all four must sum to unity. Thus, once one specifies values for the variables κ and p_e, there is only one free dimension. If $SE(\kappa)$ does not vary greatly over the range of this last "degree of freedom," then it can be approximated by a function of the first two variables. For each combination of κ and p_e, I used the following steps to generate this third dimension and to evaluate how $SE(\kappa)$ varies over it: (a) Calculate $p_o(= p_{11} + p_{22}) = \kappa(1 - p_e) + p_e$. (b) Find those p_{11} that respect the previous calculation and that keep p_{12}, p_{21}, and p_{22} as proportions. (See Appendix for details.) (c) Evaluate Equation 1 for 20 values of p_{11}, equally spaced across its allowed range, and find the mean, minimum, maximum, and standard deviation of the 20 values of the SE.

$k = 3$ and 4

The approach used for the $k = 2$ case was not feasible. Instead, I generated values of κ and p_e by using a more direct but computer intensive method. For $k = 3$, I created a series of 3 × 3 tables, each described by nine p_{ij}s, by looping over eight of the p_{ij}s. (The ninth was constrained so that all nine summed to unity.) I computed the values of κ, p_e, and $SE(\kappa)$ from each table. (To avoid low and therefore uninteresting values of κ, I constrained the quantity $p_o = \Sigma\, p_{ii}$ to equal 0.5 or more.) Those tables that gave rise to approximately the same values of κ and p_e (i.e., to $[\kappa, p_e]$ values falling within a 0.05 × 0.05 cell in the two-dimensional $[\kappa, p_e]$ grid) were grouped together, and the mean, minimum, maximum, and standard deviation of their $SE(\kappa)$s were computed. (The looping was weighted to yield a sufficiently large number of tables in each cell.) I followed a similar procedure for the case of $k = 4$.

Results

$k = 2$

Table 2 represents the way in which the values of $\sqrt{N} \cdot SE(\kappa)$ vary with κ and p_e, as well as how they vary within the (hidden)

third dimension at each (κ, p_e) point. There is more than a fourfold variation in $\sqrt{N} \cdot SE$ over the range of interest. This is made up of (a) a slightly less than twofold variation in $\sqrt{N} \cdot SE$ over the range of p_e (holding κ fixed), with $\sqrt{N} \cdot SE$ increasing consistently with increasing p_e, and (b) a slightly more than twofold variation in $\sqrt{N} \cdot SE$ over the range of κ (holding p_e fixed) but in a more quadratic fashion, with lower values of $\sqrt{N} \cdot SE$ at the two κ extremes. (Only the upper extreme is shown in Table 2.) If one holds κ and p_e fixed and if κ is large, the variation in $\sqrt{N} \cdot SE$ over the third dimension is much smaller, rarely more than 10% (on average), seldom over 5%, and commonly under 1%; that is, the error in using an average SE is between $1/\sqrt{N}$ and $10/\sqrt{N}$ percent. In other words, the magnitude of $SE(\kappa)$ is adequately predicted simply from knowledge of κ, p_e, and, of course, N. Figure 1 represents this numerical relation between $\sqrt{N} \cdot SE(\kappa)$ and κ and p_e in a more expanded and useful smoothed nomogram form.

k = 3 and 4

The variation of $SE(\kappa)$ has a similar pattern to that for $k = 2$, with the SE again increasing with p_e and decreasing as κ ap-

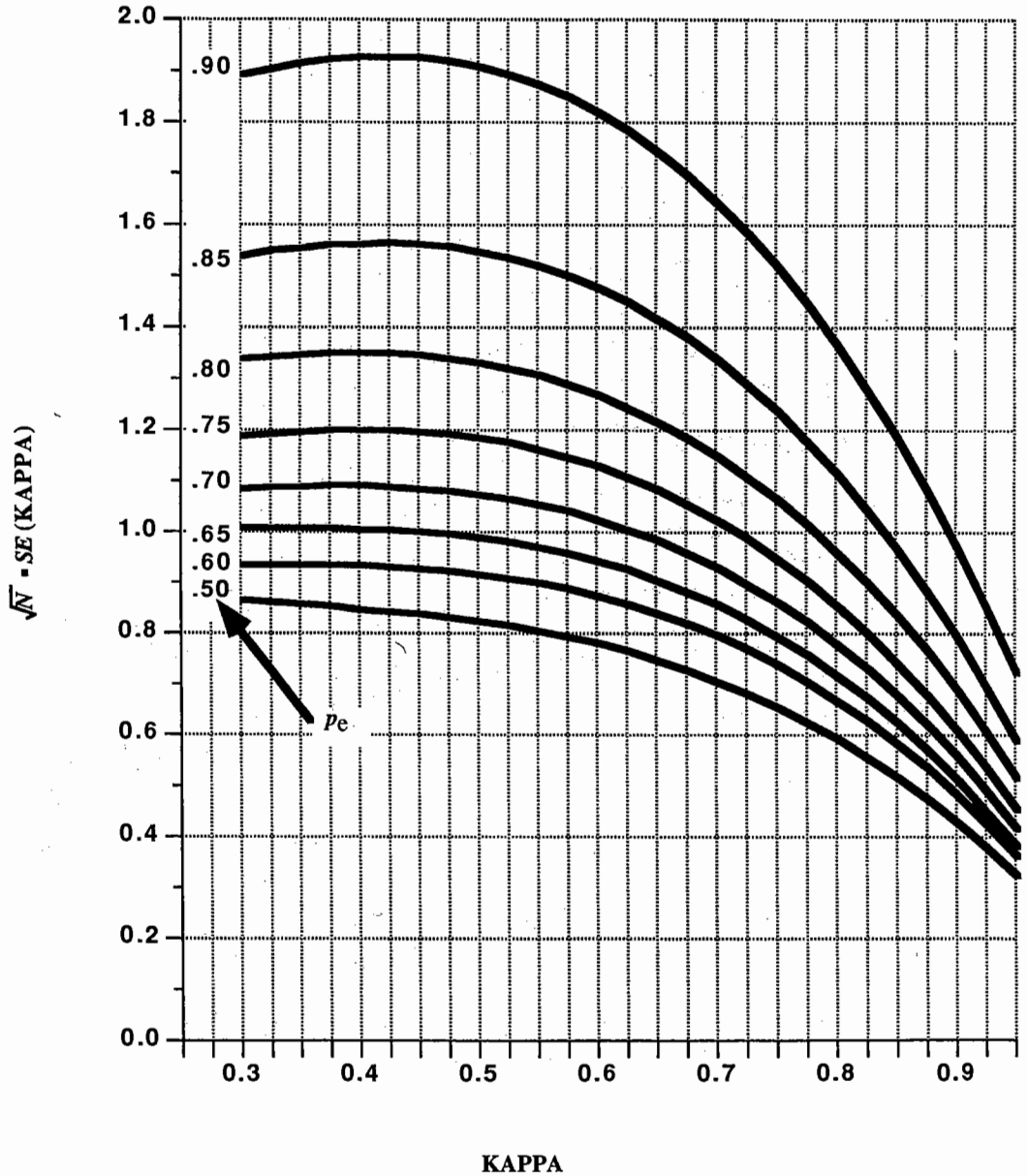

Figure 1. An approximation to $\sqrt{N} \cdot SE(\kappa)$ when κ is calculated from a 2 × 2 table. Each curve is derived from 14 data points, corresponding to $\kappa = 0.3(0.05)0.95$; each data point was obtained by averaging the SEs from 20 different tables yielding the same value of κ and p_e; the error of approximation in using this nomogram instead of Equation 1 to calculate $SE(\kappa)$ is generally on the order of $1/\sqrt{N}$ to $3/\sqrt{N}$ percent.

proaches its upper limit. As is shown in Table 3, however, the percentage error incurred in approximating each $\sqrt{N} \cdot SE(\kappa)$ by an average SE (within what is now a larger high-dimensional cell) is higher than when $k = 2$, but the approximation is still certainly adequate for planning purposes and in many cases for actual data.

Compared with $k = 2$, the cases of $k = 3$ and $k = 4$ show that (a) achievable p_e values are lower, reflecting the greater observer skill required to produce the same value of p_o, and (b) the values of SE for the same κ and p_e are lower. In the interest of readability, the smoothed relations between $\sqrt{N} \cdot SE(\kappa)$ and κ and p_e are plotted separately but on the same scale for $k = 2$, 3, and 4 in Figures 1, 2, and 3, respectively.

In Table 4, I compare the performance of the nomogram in approximating Equation 1 in several examples from textbooks and from the literature. In all but one of the cases examined, the approximation was good to two decimal places.

To illustrate the ease of use and the accuracy of the nomograms, consider Example 6 in Table 4, which is the worked example given on pages 221–222 of Fleiss (1981), where $k = 3$,

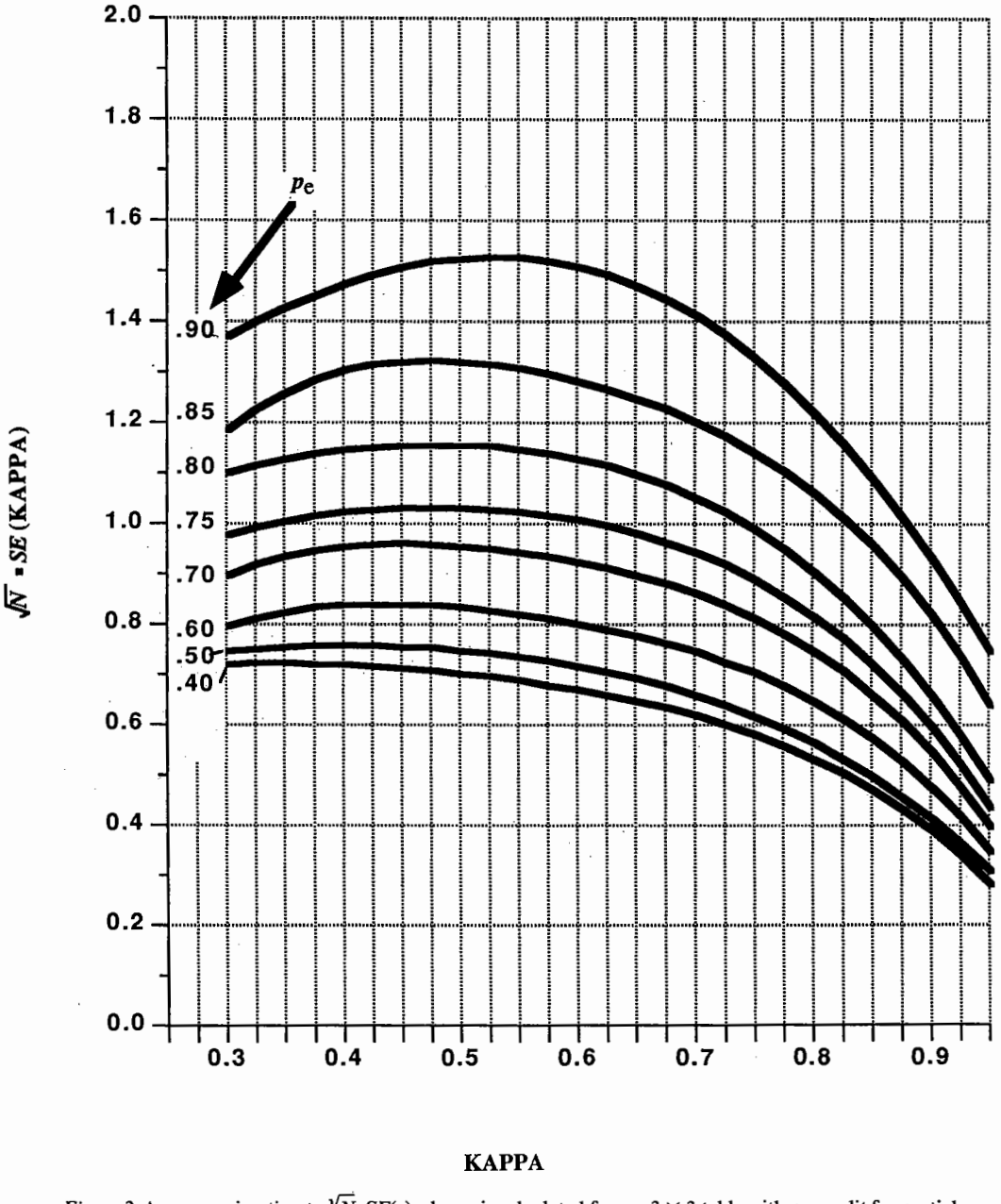

Figure 2. An approximation to $\sqrt{N} \cdot SE(\kappa)$ when κ is calculated from a 3 × 3 table with no credit for partial agreement. Each curve is derived from 14 data points, corresponding to $\kappa = 0.3(0.05)0.95$; each data point was obtained by averaging the SEs from several (>200) different tables yielding similar values of κ and p_e; the error of approximation in using this nomogram for $SE(\kappa)$ is on the order of $2/\sqrt{N}$ to $10/\sqrt{N}$ percent.

$N = 100$, $\kappa = 0.68$, and $p_e = 0.66$. By visual interpolation between the curves marked "$p_e = .60$" and "$p_e = .70$" in Figure 2, one can determine that $\sqrt{100} \cdot SE(\kappa)$ is approximately 0.84; thus, the SE itself is approximately 0.084. This compares favorably with the 0.087 calculated in the worked example.

Discussion

The purpose of this investigation was to examine the way in which $SE(\kappa)$ varies and, if possible, to simplify the expression for it. I found that it can be largely determined from κ and p_e and that the form of this relation can be represented by a useful nomogram. Before a study, the nomogram can be used to determine whether a given sample size will ensure a sufficiently small SE. Also, although this nomogram is not always perfectly accurate, it is seldom in error by more than $10/\sqrt{N}$ percent; if N is greater than 100, the error is less than 1% and, therefore, affects only the third decimal place. Thus, the nomogram can also be used with actual data if one wishes to avoid evaluating Equation 1.

The small inaccuracies in using the nomogram are inconsequential; investigators should seldom be interested in testing

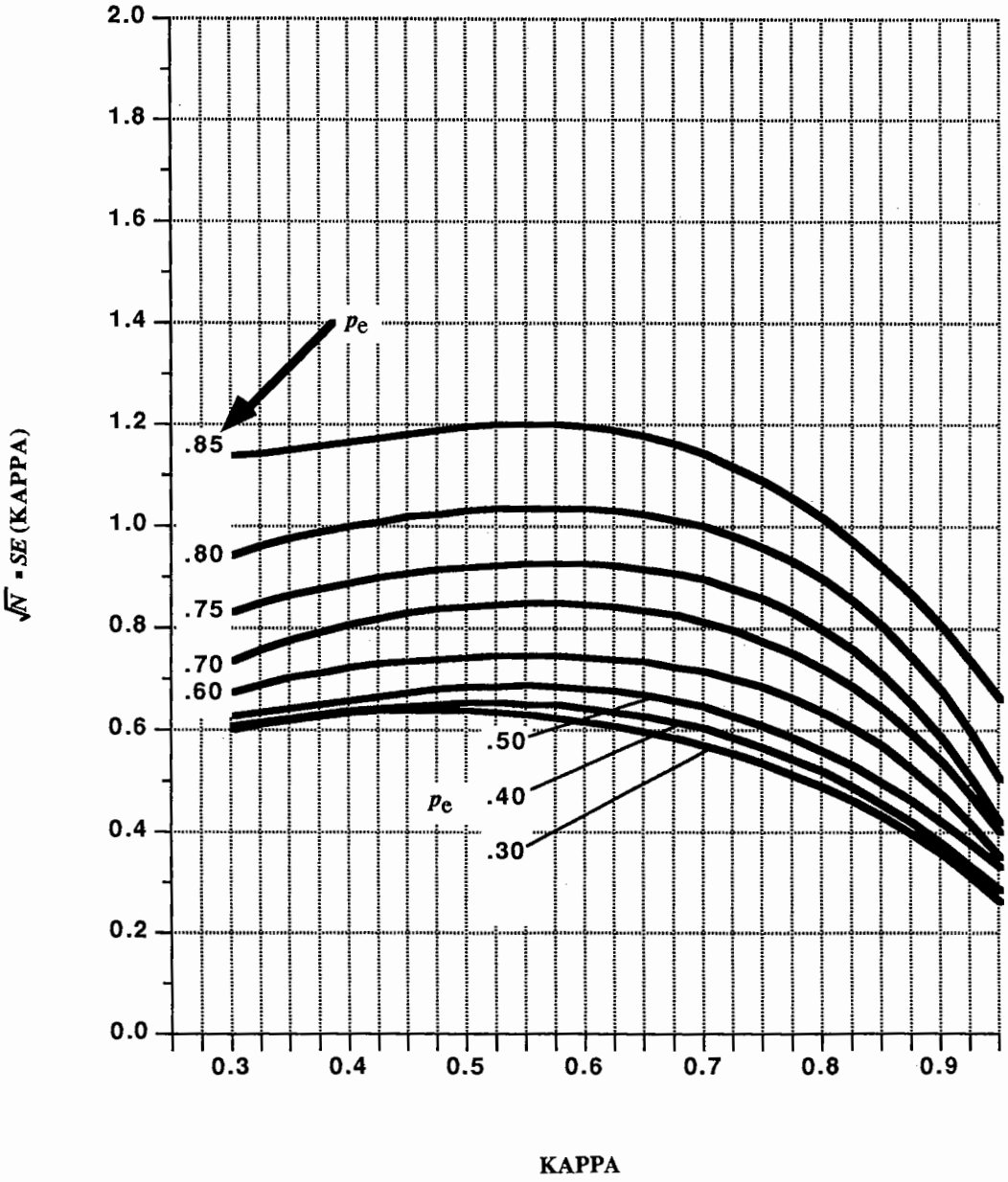

Figure 3. An approximation to $\sqrt{N} \cdot SE(\kappa)$ when κ is calculated from a 4×4 table with no credit for partial agreement. Each curve is derived from 14 data points, corresponding to $\kappa = 0.3(0.05)0.95$; each data point was obtained by averaging the SEs from several (>200) different tables yielding similar values of κ and p_e; the error of approximation in using this nomogram for $SE(\kappa)$ is on the order of $3/\sqrt{N}$ to $18/\sqrt{N}$ percent.

Table 4
Performance of the Nomogram in Approximating $\sqrt{N} \cdot SE(\kappa)$: Examples From Textbooks and From the Literature

Example	k	κ	p_e	N	*SE* calculated by using Equation 1	*SE* obtained from Figure 1, 2, or 3
1	2	.70	.66	179	0.066	0.063
2	2	.83	.67	179	0.052	0.050
3	2	.89	.54	77	0.055	0.054
4	3	.36	.35	72	0.091	0.082
5	3	.43	.48	200	0.054	0.053
6	3	.68	.66	100	0.087	0.084
7	4	.76	.49	81	0.068	0.067
8	4	.93	.70	180	0.036	0.035

whether the true κ is zero or in producing extremely precise estimates of κ; rather, they should use confidence intervals to locate and broadly categorize the true κ. What one wishes to know is whether the true κ, when expressed as a percentage, is, for example, in the 30s, 50s, 70s, or 90s. (See, e.g., the guidelines proposed by Landis & Koch, 1977.) Moreover, it is false precision to use any finer grain in a confidence interval than is achievable with these nomograms, especially because, even if one uses Equation 1, the sample sizes are seldom large enough or the underlying statistical theory accurate enough to ensure that the quoted confidence level is absolutely correct.

The pattern in the nomograms is not surprising if one examines the structure of κ. Its denominator $(1 - p_e)$ represents the fraction of the subjects in whom the two observers should be able to show their skill and training (beyond just chance agreement p_e). In other words, the effective sample size is not N but $N' = N(1 - p_e)$. Thus, a larger p_e leads to a smaller N' and a larger $SE(\kappa)$. The decreasing SE with increasing κ has a similar intuitive basis in that with N' as a denominator, κ might be thought of—very roughly—as a percentage of skill, with sample size N' and expected proportion of skill κ. If κ were to act as a binomial statistic, its SE should take the quadratic shape $[\kappa(1 - \kappa)/N']^{1/2}$. As one can discover by superimposing this function on Figures 1–3, such a binomial-based formula would underestimate $SE(\kappa)$ and be accurate only for very high κ, of 0.8 or more. Unfortunately, I have not been able to find an appropriate transformation to bring the nomograms to a full parametric form.

The above results can serve as a guide in planning agreement studies. They illustrate that one should (a) keep the number of categories (k) as large as possible to reduce p_e and $SE(\kappa)$ and (b) use categories that will not produce very skewed proportions $\{p_{i.}\}$ and $\{p_{.j}\}$, which also increase p_e. They also show that ignoring Figures 1–3 altogether and using $1/\sqrt{N}$ for $SE(\kappa)$ will often give a rough first approximation.

The entire emphasis in this article, namely, simplifying $SE(\kappa)$, presupposes that it is going to be used with the z (Gaussian) tables to form symmetric confidence intervals, tests of significance, or both. As Fleiss and Cicchetti (1978) pointed out, relatively large sample sizes are needed before, for example, 95% confidence intervals have exactly 95% coverage. (They suggested $N \geq 16k^2$.) I will report more exact confidence intervals for small samples or extreme κs in a separate article.

References

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

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

Fleiss, J. L., & Cicchetti, D. V. (1978). Inference about weighted kappa in the non-null case. *Applied Psychological Measurement, 2,* 113–117.

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

Institute for Scientific Information. (1986). This week's citation classic. *Current Contents (Social & Behavioral Sciences), 18*(3), 18.

Kraemer, H. C. (1983). Kappa coefficient. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 352–354). Toronto, Ontario, Canada: Wiley.

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

Walter, S. D. (1984). Measuring the reliability of clinical data: The case for using three observers. *Revue Epidémologie et Santé Publique,* 206–211.

Appendix

Range of p_{11} in 2×2 Tables

It can be shown that the lower and upper bounds (a, d) for p_{11} are

$$a = [p_o - (p_o^2 - 2p_o + p_e)^{1/2}]/2$$

and

$$d = [p_o + (p_o^2 - 2p_o + p_e)^{1/2}]/2.$$

When $p_e \leq 0.5$, p_{11} can vary over the entire (a, d) range. When $p_e > 0.5$, the admissible range is the union of the two intervals (a, b) and (c, d), where

$$b = [p_o - (2p_e - 1)^{1/2}]/2$$

and

$$c = [p_o + (2p_e - 1)^{1/2}]/2.$$

The three remaining variables are then calculated as

$$p_{22} = (p_o + p_e) - p_{11},$$

$$p_{21} = [(1 - p_o) + (1 + p_o^2 - 2p_e - 4p_{11}p_{22})^{1/2}]/2, \quad \text{and}$$

$$p_{12} = 1 - (p_o + p_{21}).$$

Received May 15, 1986
Revision received December 23, 1986
Accepted December 23, 1986 ■