Perspective

The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements

Purpose. This article examines and illustrates the use and interpretation of the kappa statistic in musculoskeletal research. Summary of Key Points. The reliability of clinicians' ratings is an important consideration in areas such as diagnosis and the interpretation of examination findings. Often, these ratings lie on a nominal or an ordinal scale. For such data, the kappa coefficient is an appropriate measure of reliability. Kappa is defined, in both weighted and unweighted forms, and its use is illustrated with examples from musculoskeletal research. Factors that can influence the magnitude of kappa (prevalence, bias, and nonindependent ratings) are discussed, and ways of evaluating the magnitude of an obtained kappa are considered. The issue of statistical testing of kappa is considered, including the use of confidence intervals, and appropriate sample sizes for reliability studies using kappa are tabulated. Conclusions. The article concludes with recommendations for the use and interpretation of kappa. [Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther.* 2005;85:257–268.]

Key Words: *Kappa, Measurement, Reliability, Sample size.*

Julius Sim, Chris C Wright

In musculoskeletal practice and research, there is frequently a need to determine the reliability of measurements made by clinicians—reliability here being the extent to which clinicians agree in their ratings, not merely the extent to which their ratings are associated or correlated. Defined as such, 2 types of reliability exist: (1) agreement between ratings made by 2 or more clinicians (interrater reliability) and (2) agreement between ratings made by the same clinician on 2 or more occasions (intrarater reliability).

In some cases, the ratings in question are on a continuous scale, such as joint range of motion or distance walked in 6 minutes. In other instances, however, clinicians' judgments are in relation to discrete categories. These categories may be nominal (eg, "present," "absent") or ordinal (eg, "mild," "moderate," "severe"); in each case, the categories are mutually exclusive and collectively exhaustive, so that each case falls into one, and only one, category. A number of recent studies have used such data to examine interrater or intrarater reliability in relation to: clinical diagnoses or classifications,[1–4] assessment findings,[5–9] and radiographic signs.[10–12] These data require specific statistical methods to assess reliability, and the kappa (κ) statistic is commonly used for this purpose. This article will define and illustrate the kappa coefficient and will examine some potentially problematic issues connected with its use and interpretation. Sample size requirements, which previously were not readily available in the literature, also are provided.

If used and interpreted appropriately, the kappa coefficient provides valuable information on the reliability of diagnostic and other examination procedures.

Nature and Purpose of the Kappa Statistic

A common example of a situation in which a researcher may want to assess agreement on a nominal scale is to determine the presence or absence of some disease or condition. This agreement could be determined in situations in which 2 researchers or clinicians have used the same examination tool or different tools to determine the diagnosis. One way of gauging the agreement between 2 clinicians is to calculate *overall percentage of agreement* (calculated over all paired ratings) or *effective percentage of agreement* (calculated over those paired ratings where at least one clinician diagnoses presence of the disease).[13] Although these calculations provide a measure of agreement, neither takes into account the agreement that would be expected purely by chance. If clinicians agree purely by chance, they are not really "agreeing" at all; only agreement beyond that expected by chance can be considered "true" agreement. Kappa is such a measure of "true" agreement.[14] It indicates the proportion of agreement beyond that expected by chance, that is, the *achieved* beyond-chance agreement as a proportion of the *possible* beyond-chance agreement.[15] It takes the form:

$$(1) \qquad \kappa = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}}$$

In terms of symbols, this is:

$$(2) \qquad \kappa = \frac{P_o - P_c}{1 - P_c}$$

where P_o is the proportion of observed agreements and P_c is the proportion of agreements expected by chance. The simplest use of kappa is for the situation in which 2 clinicians each provide a single rating of the same patient, or where a clinician provides 2 ratings of the same patient, representing interrater and intrarater reliability, respectively. Kappa also can be adapted for more than one rating per patient from each of 2 clinicians,[16,17] or for situations where more than 2 clinicians rate each patient or where each clinician may not rate every patient.[18] In this article, however, our focus will be on the simple situation where 2 raters give an independent single rating for each patient or where a single rater provides 2 ratings for each patient. Here, the concern is with how well these ratings agree, not with their relationship with some "gold standard" or "true" diagnosis.[19]

The data for paired ratings on a 2-category nominal scale are usually displayed in a 2 × 2 contingency table, with the notation indicated in Table 1.[20] This table shows data from 2 clinicians who assessed 39 patients in relation to the relevance of lateral shift, according to the McKenzie method of low back pain assessment.[9] Cells a and d indicate, respectively, the numbers of patients for whom both clinicians agree on the relevance or nonrelevance of lateral shift. Cells b and c indicate the numbers of

J Sim, PhD, is Professor, Primary Care Sciences Research Centre, Keele University, Keele, Staffordshire ST5 5BG, United Kingdom (j.sim@keele.ac.uk). Address all correspondence to Dr Sim.

CC Wright, BSc, is Principal Lecturer, School of Health and Social Sciences, Coventry University, Coventry, United Kingdom.

Both authors provided concept/idea and writing.

Table 1.
Diagnostic Assessments of Relevance of Lateral Shift by 2 Clinicians, From Kilpikoski et al[9] (κ=.67)[a]

		Clinician 2		
		Relevant	Not relevant	Total
Clinician 1	Relevant	*a* 22	*b* 2	g_1 24
	Not relevant	*c* 4	*d* 11	g_2 15
	Total	f_1 26	f_2 13	*n* 39

[a] The letters in the upper left-hand corners of the cells indicate the notation used for a 2 × 2 contingency table. The main diagonal cells (*a* and *d*) represent agreement, and the off-diagonal cells (*b* and *c*) represent disagreement.

patients on whom the clinicians disagree. For clinician 2, the total numbers of patients in whom lateral shift was deemed relevant or not relevant are given in the marginal totals, f_1 and f_2, respectively. The corresponding marginal totals for clinician 1 are g_1 and g_2.

Summing the frequencies in the main diagonal cells (cells *a* and *d*) gives the frequency of observed agreement. Dividing by *n* gives the proportion of observed agreement. Thus, the proportion of observed agreement in Table 1 is:

$$(3) \qquad P_o = \frac{(a+d)}{n} = \frac{22+11}{39} = .8462$$

The proportion of expected agreement is based on the assumption that assessments are independent between clinicians. Therefore, the frequency of chance agreement for relevance and nonrelevance of lateral shift is calculated by multiplying the marginal totals corresponding to each cell on the main diagonal and dividing by *n*. Summing across chance agreement in these cells and dividing by *n* gives the proportion of expected agreement. For the data in Table 1, this is:

$$(4)$$

$$P_c = \frac{\left(\frac{f_1 \times g_1}{n}\right) + \left(\frac{f_2 \times g_2}{n}\right)}{n} = \frac{\left(\frac{26 \times 24}{39}\right) + \left(\frac{13 \times 15}{39}\right)}{39}$$

$$= \frac{16+5}{39} = .5385$$

Substituting into the formula:

$$(5) \qquad \kappa = \frac{P_o - P_c}{1 - P_c} = \frac{.8462 - .5385}{1 - .5385} = .67$$

The range of possible values of kappa is from –1 to 1, though it usually falls between 0 and 1. Unity represents perfect agreement, indicating that the raters agree in their classification of every case. Zero indicates agreement no better than that expected by chance, as if the raters had simply "guessed" every rating. A negative kappa would indicate agreement worse than that expected by chance.[21] However, this rarely occurs in clinical contexts, and, when it does, the magnitude of the negative coefficient is usually small (theoretically a value of -1 can be attained if 2 raters are being considered, though with more than 2 raters the possible minimum value will be higher).[22]

The kappa coefficient does not itself indicate whether disagreement is due to random differences (ie, those due to chance) or systematic differences (ie, those due to a consistent pattern) between the clinicians' ratings,[23] and the data should be examined accordingly. The Figure shows the relationship of kappa to overall and chance agreement schematically.[24]

Adaptations of the Kappa Coefficient

The kappa coefficient can be used for scales with more than 2 categories. Richards et al[12] assessed intraobserver and interobserver agreement of radiographic classification of scoliosis in relation to the King classification system. The King system is a multicategory nominal scale by means of which radiographs of the spine can be classified into 1 of 5 types of spinal curve. However, many multicategory scales are ordinal, and in such cases it is important to retain the hierarchical nature of the categories.

Table 2 presents the results of a hypothetical reliability study of assessments of movement-related pain, on 2 occasions by a single examiner, during which time pain would not have been expected to change. The assessment categories were "no pain," "mild pain," "moderate pain," and "severe pain." These categories are clearly ordinal, in that they reflect increasing levels of movement-related pain. Here, disagreement by 1 scale point (eg, "no pain"–"mild pain") is less serious than disagreement by 2 scale points (eg, "no pain"–"moderate pain"). To reflect the degree of disagreement, kappa can be weighted, so that it attaches greater emphasis to large differences between ratings than to small differences. A number of methods of weighting are available,[25] but quadratic weighting is common (Appendix). Weighted kappa penalizes disagreements in terms of their seriousness, whereas unweighted kappa treats all disagreements equally. Unweighted kappa, therefore, is inappropriate for ordinal scales.[26] Because in this example most disagreements are of only a single category, the quadratic weighted kappa (.67) is higher than the unweighted kappa (.55). Different weighting schemes will produce different values of weighted kappa on the same data; for example, linear weighting gives a kappa of .61 for the data in Table 2.

Such weightings also can be applied to a nominal scale with 3 or more categories, if certain disagreements are considered more serious than others. Table 3 shows data for the agreement between 2 raters on the presence of a derangement, dysfunction, or postural syndrome, in terms of the classification of spinal pain originally proposed by McKenzie.[27] The value of kappa for these data is .46. Normally, in the calculation of kappa, the agreement cells (cells *a*, *e*, and *i*) would be given a weighting of unity, and the remaining disagreement cells would be given a weighting of zero (Appendix). If it were felt, however, that a disagreement between a dysfunctional syndrome and a postural syndrome is of less concern clinically than a disagreement between a derangement syndrome and a dysfunctional syndrome, or between a derangement syndrome and a postural syndrome, this could be represented by applying a linear weighting to the cell frequencies. Accordingly, cells *h* and *f* would have a weight of .5, while the weights for cells *b*, *c*, *d*, and *g* would remain at zero. With this weighting, the value of kappa becomes .50. Because 16 disagreements (cells *h* and *f*) of the total of 36 disagreements are now treated as less serious through the linear weighting, kappa has increased.

For a nominal scale with more than 2 categories, the obtained value of kappa does not identify individual categories on which there may be either high or low agreement.[28] The use of weighting also may serve to determine the sources of disagreement between raters on a nominal scale with more than 2 categories and the effect of these disagreements on the values of kappa.[29] A cell representing a particular disagreement can be assigned a weight

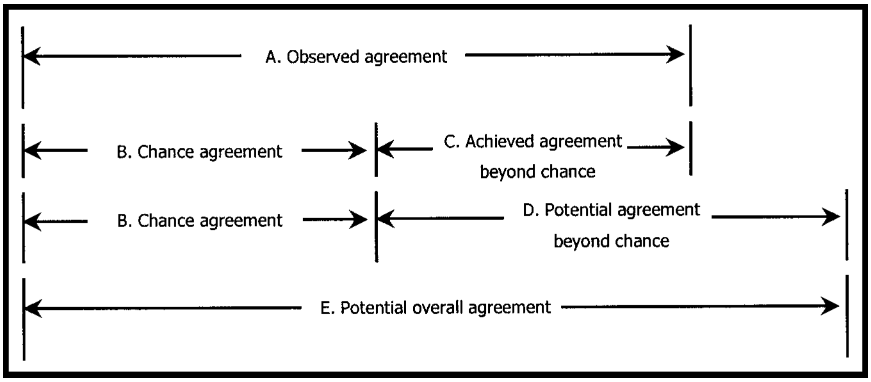
A. Observed agreement
B. Chance agreement
C. Achieved agreement beyond chance
B. Chance agreement
D. Potential agreement beyond chance
E. Potential overall agreement

Figure.
Schematic representation of the relationship of kappa to overall and chance agreement. Kappa=C/D. Adapted from Rigby.[24]

Table 2.
Test-Retest Agreement of Ratings of Movement-Related Pain at the Shoulder Joint (Hypothetical Data)[a]

		Test 2				Total
		No pain	Mild pain	Moderate pain	Severe pain	
Test 1	No pain	15 (1) [1]	3 (.67) [.89]	1 (.33) [.56]	1 (0) [0]	20
	Mild pain	4 (.67) [.89]	18 (1) [1]	3 (.67) [.89]	2 (.33) [.56]	27
	Moderate pain	4 (.33) [.56]	5 (.67) [.89]	16 (1) [1]	4 (.67) [.89]	29
	Severe pain	1 (0) [0]	2 (.33) [.56]	4 (.67) [.89]	17 (1) [1]	24
Total		24	28	24	24	100

[a] Figures in parentheses are linear kappa weights; figures in brackets are quadratic kappa weights. Unweighted κ=.55; linear weighted κ=.61; quadratic weighted κ=.67.

Table 3.
Interrater Agreement of Ratings of Spinal Pain (Hypothetical Data)[a]

		Clinician 2			Total
		Derangement syndrome	Dysfunctional syndrome	Postural syndrome	
Clinician 1	Derangement syndrome	*a* 22	*b* 10	*c* 2	34
	Dysfunctional syndrome	*d* 6	*e* 27	*f* 11	44
	Postural syndrome	*g* 2	*h* 5	*i* 17	24
Total		30	42	30	102

[a] Unweighted κ=.46; cells *b* and *d* weighted as agreement κ=.50; cells *f* and *h* weighted as agreement κ=.55.

Table 4.
(A) Assessment of the Presence of Lateral Shift, From Kilpikoski et al[9] (κ=.18); (B) the Same Data Adjusted to Give Equal Agreements in Cells *a* and *d*, and Thus a Low Prevalence Index (κ=.54)

A		Clinician 2		Total
		Present	Absent	
Clinician 1	Present	*a* 28	*b* 3	31
	Absent	*c* 6	*d* 2	8
Total		34	5	39

B		Clinician 2		Total
		Present	Absent	
Clinician 1	Present	*a* 15	*b* 3	18
	Absent	*c* 6	*d* 15	21
		21	18	39

representing agreement (unity), effectively treating this source of disagreement as an agreement, while leaving unchanged the weights for remaining sources of disagreement. The alteration that this produces in the value of kappa serves to quantify the effect of the identified disagreement on the overall agreement, and all possible sources of disagreement can be compared in this way. Returning to the data in Table 3, if we weight as agreements those instances in which the raters disagreed between derangement and dysfunction syndromes (cells *b* and *d*), kappa rises from .46 without weighting to .50 with weighting. If alternatively we apply agreement weighting to disagreements between dysfunctional and postural syndromes (cells *f* and *h*), kappa rises more markedly to .55. As the disagreement between dysfunctional and postural syndromes produces the greater increase in kappa, it can be seen to contribute more to the overall disagreement than that between derangement and dysfunctional syndromes. This finding might indicate that differences between postural and dysfunctional syndromes are more difficult to determine than differences between derangement and dysfunctional syndromes. This information might lead to retraining of the raters or rewording of examination protocols.

In theory, kappa can be applied to ordinal categories derived from continuous data. For example, joint ranges of motion, measured in degrees, could be placed into 4 categories: "unrestricted," "slightly restricted," "moderately restricted," and "highly restricted." However, the results from such an analysis will depend largely on the choice of the category limits. As this choice is in many cases arbitrary, the value of kappa produced may have little meaning. Furthermore, this procedure involves needless sacrifice of information in the original scale and will normally give rise to a loss of statistical power.[30,31] It is far preferable to analyze the reliability of data obtained with the original continuous scale[32] using other methods such as the intraclass correlation coefficient,[33] the standard error of measurement,[34] or the bias and limits of agreement.[35]

Determinants of the Magnitude of Kappa

As previously noted, the magnitude of the kappa coefficient represents the proportion of agreement greater than that expected by chance. The interpretation of the coefficient, however, is not so straightforward, as there are other factors that can influence the magnitude of the coefficient or the interpretation that can be placed on a given magnitude. Among those factors that can influence the magnitude of kappa are prevalence, bias, and nonindependence of ratings.

Prevalence

The kappa coefficient is influenced by the prevalence of the attribute (eg, a disease or clinical sign). For a situation in which raters choose between classifying cases as either positive or negative in respect to such an attribute, a prevalence effect exists when the proportion of agreements on the positive classification differs from that of the negative classification. This can be expressed by the *prevalence index*. Using the notation from Table 1, this is:

$$\text{prevalence index}=\frac{|a-d|}{n} \tag{6}$$

where $|a-d|$ is the absolute value of the difference between the frequencies of these cells (ie, ignoring the sign) and n is the number of paired ratings.

If the prevalence index is high (ie, the prevalence of a positive rating is either very high or very low), chance agreement is also high and kappa is reduced accordingly.[29] This can be shown by considering further data from Kilpikoski et al[9] on the presence or absence of lateral shift (Tab. 4). In Table 4A, the prevalence index is high:

$$\text{prevalence index}=\frac{|28-2|}{39}=.67 \tag{7}$$

The proportion of chance agreement, therefore, also is relatively high (.72), and the value of kappa is .18. In

Table 5.
(A) Contingency Table Showing Nearly Symmetrical Disagreements in Cells *b* and *c*, and Thus a Low Bias Index (κ=.12); (B) Contingency Table With Asymmetrical Disagreements in Cells *b* and *c*, and Thus a Higher Bias Index (κ=.20)[a]

A		Clinician 2		Total	B		Clinician 2		Total
		Present	Absent				Present	Absent	
Clinician 1	Present	*a* 29	*b* 21	50	Clinician 1	Present	*a* 29	*b* 6	35
	Absent	*c* 23	*d* 27	50		Absent	*c* 38	*d* 27	65
Total		52	48	*n* 100			67	33	*n* 100

[a] Hypothetical data for diagnoses of spondylolisthesis ("present" or "absent") by 2 clinicians.

Table 4B, however, there is a lower prevalence index of zero. Although the raters agree on the same number of cases (30) as in Table 4A, the low prevalence index reduces chance agreement to .50, and the value of kappa accordingly rises to .54. From Table 4B, prevalence index=$|15-15|/39=0$, $P_o=(28+2)/39=.7692$, $P_c=[(21\times18)/39+(18\times21)/39]/39=.4970$. Thus,

$$(8)\qquad \kappa=(P_o-P_c)/(1-P_c)=(.7692-.4970)/(1-.4970)=.2695/.5030=.54$$

This illustrates the first of 2 paradoxes[20]: when there is a large prevalence index, kappa is lower than when the prevalence index is low or zero. The effect of prevalence on kappa is greater for large values of kappa than for small values.[36]

Bannerjee and Fielding[37] suggest that it is the *true* prevalence in the population that affects the magnitude of kappa. This is not wholly accurate, as the prevalence index does not provide a direct indication of the true prevalence of the disease. Rather, if a disease is either very common or very rare, this will predispose clinicians to diagnose or not to diagnose it, respectively, so that the prevalence index provides only an indirect indication of true prevalence, mediated by the clinicians' diagnostic behavior.

Because the magnitude of kappa is affected by the prevalence of the attribute, kappa on its own is difficult to interpret meaningfully unless the prevalence index is taken into account.

Bias

Bias is the extent to which the raters disagree on the proportion of positive (or negative) cases and is reflected in a difference between cells *b* and *c* in Table 1. The bias index is:

$$(9)\qquad \text{bias index}=\frac{|b-c|}{n}$$

Bias affects our interpretation of the magnitude of the coefficient. Table 5 shows hypothetical data for 2 clinicians' diagnosis of spondylolisthesis in 100 patients. In both Table 5A and Table 5B, the proportion of cases on which the raters agree is the same, at .56, but the pattern of disagreements differs between the 2 tables because each clinician rates a differing proportion of cases as positive. In Table 5A, the proportions of cases rated as positive are .50 and .52 for clinicians 1 and 2, respectively, whereas the corresponding proportions in Table 5B are .35 and .67. In Table 5A, disagreement is close to symmetrical. The bias index is accordingly low:

$$(10)\qquad \text{bias index}=\frac{|23-21|}{100}=.02$$

In contrast, in Table 5B the disagreements are asymmetrical. There is, therefore, a much higher bias index in Table 5B:

$$(11)\qquad \text{bias index}=\frac{|38-6|}{100}=.32$$

Owing to the much greater bias in Table 5B than in Table 5A, the resulting kappa coefficients are different (.20 and .12, respectively). This gives rise to the second paradox[20]: when there is a large bias, kappa is higher than when bias is low or absent. In contrast to prevalence, the effect of bias is greater when kappa is small than when it is large.[36] Just as with prevalence, the magnitude of kappa should be interpreted in the light of the bias index.

Nonindependent Ratings

An important assumption underlying the use of the kappa coefficient is that errors associated with clinicians'

Table 6.
(A) Data Reported by Kilpikoski et al[9] for Judgments of Directional Preference by 2 Clinicians (κ=.54); (B) Cell Frequencies Adjusted to Minimize Prevalence and Bias Effects, Giving a Prevalence-Adjusted Bias-Adjusted κ of .79

A		Clinician 2		Total
		Present	Absent	
Clinician 1	Present	*a* 32	*b* 1	33
	Absent	*c* 3	*d* 3	6
Total		35	4	39

B		Clinician 2		Total
		Present	Absent	
Clinician 1	Present	*a* 18	*b* 2	20
	Absent	*c* 2	*d* 17	19
		20	19	39

ratings are independent.[38–40] This requires the patients or subjects to be independent (so that any individual can contribute only one paired rating) and ratings to be independent (so that each observer should generate a rating without knowledge, and thus without influence, of the other observer's rating).[40] The fact that ratings are related in the sense of pertaining to the same case, however, does not contravene the assumption of independence.

The kappa coefficient, therefore, is not appropriate for a situation in which one observer is required to either confirm or disconfirm a known previous rating from another observer. In such a situation, agreement on the underlying attribute is contaminated by agreement on the *assessment* of that attribute, and the magnitude of kappa is liable to be inflated. Equally, as with all measures of intratester reliability, ratings on the first testing may sometimes influence those given on the second occasion, which will threaten the assumption of independence. In this way, apparent agreement may reflect more a recollection of the previous decision than a genuine judgment as to the appropriate classification. In a situation in which the clinician is doubtful as to the appropriate classification, this recollection may sway the decision in favor of agreement rather than disagreement with the previous decision. Thus, "agreements" that represent a decision to classify in the same way will be added to agreements on the actual attribute. This will tend to increase the value of kappa.[38]

Accordingly, studies of either interrater or intrarater reliability should be designed in such a way that ratings are, as far as possible, independent, otherwise kappa values may be inappropriately inflated. Equally, where a study appears not to have preserved independence between ratings, kappa should be interpreted cautiously. Strictly, there will always be some degree of dependence between ratings in an intrarater study.[38] Various strategies can be used, however, to minimize this dependence. The time interval between repeat ratings is important. If the interval is too short, the rater might remember the previously recorded rating; if the interval is too long, then the attribute under examination might have changed. Streiner and Norman[33] stated that an interval of 2 to 14 days is usual, but this will depend on the attribute being measured. Stability of the attribute being rated is crucial to the period between repeated ratings. Thus, trait attributes pose fewer problems for intrarater assessment (because longer periods of time may be left between ratings) than state attributes, which are more labile. Some suggestions to overcome the bias due to memory include: having as long a time period as possible between repeat examinations, blinding raters to their first rating (although this might be easier with numerical data than with diagnostic categories), and different random ordering of patients or subjects on each rating occasion and for each rater.

Adjusting Kappa

Because both prevalence and bias play a part in determining the magnitude of the kappa coefficient, some statisticians have devised adjustments to take account of these influences.[36] Kappa can be adjusted for high or low prevalence by computing the average of cells *a* and *d* and substituting this value for the actual values in those cells. Similarly, an adjustment for bias is achieved by substituting the mean of cells *b* and *c* for those actual cell values. The kappa coefficient that results is referred to as *PABAK* (prevalence-adjusted bias-adjusted kappa). Table 6A shows data from Kilpikoski et al[9] for assessments of directional preference (ie, the direction of movement that reduces or abolishes pain) in patients evaluated according to the McKenzie system; kappa for these data is .54. When the cell frequencies are adjusted to minimize prevalence and bias, this gives the cell values shown in Table 6B, with a PABAK of .79.

Hoehler[41] is critical of the use of PABAK because he believes that the effects of bias and prevalence on the magnitude of kappa are themselves informative and should not be adjusted for and thereby disregarded. Thus, the PABAK could be considered to generate a value for kappa that does not relate to the situation in which the original ratings were made. Table 6B represents very different diagnostic behavior from Table 6A, as indicated by the change in the marginal totals. Furthermore, all of the frequencies in the cells have changed between Table 6A and Table 6B.

Therefore, the PABAK coefficient on its own is uninformative because it relates to a hypothetical situation in which no prevalence or bias effects are present. However, if PABAK is presented in addition to, rather than in place of, the obtained value of kappa, its use may be considered appropriate because it gives an indication of the likely effects of prevalence and bias alongside the true value of kappa derived from the specific measurement context studied. Cicchetti and Feinstein[42] argued, in a similar vein to Hoehler,[41] that the effects of the prevalence and bias "penalize" the value of kappa in an appropriate manner. However, they also stated that a single "omnibus" value of kappa is difficult to interpret, especially when trying to diagnose the possible cause of an apparent lack of agreement. Byrt et al[36] recommended that the prevalence index and bias index should be given alongside kappa, and other authors[42,43] have suggested that the separate proportions of positive and negative agreements should be quoted as a means of alerting the reader to the possibility of prevalence or bias effects. Similarly, Gjørup[44] suggested that kappa values should be accompanied by the original data in a contingency table.

Interpreting the Magnitude of Kappa

Landis and Koch[45] have proposed the following as standards for strength of agreement for the kappa coefficient: ≤0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect. Similar formulations exist,[46–48] but with slightly different descriptors. The choice of such benchmarks, however, is inevitably arbitrary,[29,49] and the effects of prevalence and bias on kappa must be considered when judging its magnitude. In addition, the magnitude of kappa is influenced by factors such as the weighting applied and the number of categories in the measurement scale.[32,49–51] When weighted kappa is used, the choice of weighting scheme will affect its magnitude (Appendix). The larger the number of scale categories, the greater the potential for disagreement, with the result that unweighted kappa will be lower with many categories than with few.[32] If quadratic weighting is used, however, kappa increases with the number of categories, and this is most marked in the range from 2 to 5 categories.[50] For linear weighting, kappa varies much less with the number of categories than for quadratic weighting, and may increase or decrease with the number of categories, depending on the distribution of the underlying trait.[50] Caution, therefore, should be exercised when comparing the magnitude of kappa across variables that have different prevalence or bias or that are measured on dissimilar scales or across situations in which different weighting schemes have been applied to kappa.

Table 7.
Data on Assessments of Stiffness at C1–2 From Smedmark et al[7,a]

		Clinician 2		Total
		Stiffness	No stiffness	
Clinician 1	Stiffness	2 (3)	1 (0)	3
	No stiffness	7 (6)	50 (51)	57
Total		9	51	60

[a] Figures in the cells represent the observed ratings; those in parentheses are the ratings that would secure maximum agreement given the marginal totals. Observed κ=.28; maximum attainable κ=.46. Smedmark et al did not specify the distribution of the 8 disagreements across the off-diagonal cells, but the figures in the table correspond to their reported κ.

Dunn[49] suggested that interpretation of kappa is assisted by also reporting the maximum value it could attain for the set of data concerned. To calculate the maximum attainable kappa ($\kappa_{\max}$), the proportions of positive and negative judgments by each clinician (ie, the marginal totals) are taken as fixed, and the distribution of paired ratings (ie, the cell frequencies *a*, *b*, *c*, and *d* in Tab. 1) is then adjusted so as to represent the greatest possible agreement. Table 7 illustrates this process, using data from a study of therapists' examination of passive cervical intervertebral motion.[7] Ratings were on a 2-point scale (ie, "stiffness"/"no stiffness"). Table 7 shows that clinician 1 judged stiffness to be present in 3 subjects, whereas clinician 2 arrived at a figure of 9. Thus, the maximum possible agreement on stiffness is limited to 3 subjects, rather than the actual figure of 2. Similarly, clinician 1 judged 57 subjects to have no stiffness, compared with 51 subjects judged by clinician 2 to have no stiffness; therefore, for "no stiffness," the maximum agreement possible is 51 subjects, rather than 50. That is, the maximum possible agreement for either presence or absence of the disease is the smaller of the marginal totals in each case. The remaining 6 ratings (60 – [3 + 51] = 6) are allocated to the cells that represent disagreement, in order to maintain the marginal total; thus, these ratings are allocated to cell *c*. For these data, $\kappa_{\max}$ is .46, compared with a kappa of .28.

In contrast to the PABAK, $\kappa_{\max}$ serves to gauge the strength of agreement while preserving the proportions of positive ratings demonstrated by each clinician. In effect, it provides a reference value for kappa that preserves the individual clinician's overall propensity to diagnose a condition or select a rating (within the restraints imposed by the marginal totals; f_1, f_2, g_1, and g_2 in Tab. 1). In some situations, equal marginal totals are not necessarily to be anticipated, owing to recognized pre-existing biases,[52] such as when comparing observers with different levels of experience, tools or assessment protocols with inherently different sensitivity,[53] or examiners who can be expected to utilize classification criteria of different stringency.[14] These a priori sources of disagreement will give rise to different marginal totals and will, in turn, be reflected in $\kappa_{\max}$.

For a given reliability study, the difference between kappa and 1 indicates the total *unachieved* agreement beyond chance. The difference between kappa and $\kappa_{\max}$, however, indicates the unachieved agreement beyond chance, within the constraints of the marginal totals. Accordingly, the difference between $\kappa_{\max}$ and 1 shows the effect on agreement of imbalance in the marginal totals. Thus, $\kappa_{\max}$ reflects the extent to which the raters' ability to agree is constrained by pre-existing factors that tend to produce unequal marginal totals, such as differences in their diagnostic propensities or dissimilar sensitivity in the tools they are using. This provides useful information.

Statistical Significance

The kappa coefficient does not reflect sampling error, and where it is intended to generalize the findings of a reliability study to a population of raters, the coefficient is frequently assessed for statistical significance through a hypothesis test. A 1-tailed test is often considered appropriate when the null hypothesis states a value of zero for kappa because a negative value of kappa does not normally have a meaningful interpretation.[29]

In a practical situation in which ratings are compared across clinicians, agreement will usually be better than that expected by chance, and specifying a zero value for kappa in the null hypothesis is therefore not very meaningful.[54,55] Thus, the value in the null hypothesis should usually be set at a higher level (eg, to determine whether the population value of kappa is greater than .40, on the basis that any value lower than this might be considered clinically "unacceptable"). The minimum acceptable value of kappa will depend on the clinical context. For example, the distress and risk assessment method (DRAM)[56] can be used to identify patients with low back pain who are at risk for poor outcome. If this categorization determines whether or not such patients are enrolled in a management program, the minimal acceptable level of agreement on this categorization might be set fairly high, so that decisions on patient management are made with a high degree of consistency. If agreement on the test were particularly important and the kappa were lower than acceptable, it might mean that clinicians need more training in the testing technique or the protocol needs to be reworded. When a value greater than zero is specified for kappa in the null hypothesis, a 2-tailed test is preferable to a 1-tailed test. This is because there is no theoretical reason to assume that the reliability of a test's results or a diagnosis will necessarily be superior to a stated threshold for clinical importance. One-tailed tests should be reserved for occasions when testing a null hypothesis that kappa is zero.

The statistical hypothesis test provides only binary information: Is there sufficient evidence that the population value of kappa is greater than .40, or not? A more useful approach is to construct a confidence interval around the sample estimate of kappa, using the standard error of kappa (Appendix) and the *z* score corresponding to the desired level of confidence.[57] The confidence interval will indicate a range of plausible values for the "true" value of kappa, with a stated level of confidence. As with hypothesis testing, however, it may make sense to evaluate the lower limit of the confidence interval against a clinically meaningful minimum magnitude, such as .40, rather than against a zero value. Taking the data in Table 5A, for the obtained kappa of .54, the 2-sided 95% confidence interval is given by:

$$(12) \quad .54-(1.96\times.199) \text{ to } .54+(1.96\times.199)=.15 \text{ to } .93$$

where 1.96 is the *z* score corresponding to a 95% 2-sided confidence level and .199 is the standard error of the obtained kappa. Therefore, although the null hypothesis that kappa is no greater than zero can be rejected (because zero lies below the lower confidence limit), the null hypothesis that kappa is equal to .40 is retained (because .40 lies within the confidence interval).

Prior to undertaking a reliability study, a sample size calculation should be performed so that a study has a stated probability of detecting a statistically significant kappa coefficient or of providing a confidence interval of a desired width.[58,59] Table 8 gives the minimum number of participants required to detect a kappa coefficient as statistically significant, with various values of the proportion of positive ratings made on a dichotomous variable by 2 raters, specifically, $(f_1 + g_1)/2n$ in terms of the notation in Table 1. A number of points should be noted in relation to this table. First, the sample sizes given assume no bias between raters. Second, except where the value of kappa stated in the null hypothesis is zero, sample size requirements are greatest when the proportion of positive ratings is either high or low. Third, given that the minimum value of kappa deemed to be clinically important will depend on the measurement context, in addition to a null value of zero, non-zero null values between .40 and .70 have been included in Table 8. Finally, following earlier comments on 1- and 2-tailed tests, the figures given are for 2-tailed tests at a significance level of .05, except where the value of kappa in the null hypothesis is zero, when figures for a 1-tailed test also are given.

Table 8.
The Number of Subjects Required in a 2-Rater Study to Detect a Statistically Significant κ ($P \leq .05$) on a Dichotomous Variable, With Either 80% or 90% Power, at Various Proportions of Positive Diagnoses, and Assuming the Null Hypothesis Value of Kappa to be .00, .40, .50, .60, or .70[a]

		1-Tailed Test Null Value=.00		2-Tailed Test Null Value=.00		2-Tailed Test Null Value=.40		2-Tailed Test Null Value=.50		2-Tailed Test Null Value=.60		2-Tailed Test Null Value=.70	
Proportion of Positive Ratings	Kappa to Detect	*n* at 80% Power	*n* at 90% Power	*n* at 80% Power	*n* at 90% Power	*n* at 80% Power	*n* at 90% Power	*n* at 80% Power	*n* at 90% Power	*n* at 80% Power	*n* at 90% Power	*n* at 80% Power	*n* at 90% Power
.10	.40	39	54	50	66								
.30	.40	39	54	50	66								
.50	.40	39	54	50	66								
.70	.40	39	54	50	66								
.90	.40	39	54	50	66								
.10	.50	25	35	32	43	1,617	2,164						
.30	.50	25	35	32	43	762	1,020						
.50	.50	25	35	32	43	660	883						
.70	.50	25	35	32	43	762	1,020						
.90	.50	25	35	32	43	1,617	2,164						
.10	.60	18	24	22	30	405	541	1,519	2,034				
.30	.60	18	24	22	30	191	255	689	922				
.50	.60	18	24	22	30	165	221	589	789				
.70	.60	18	24	22	30	191	255	689	922				
.90	.60	18	24	22	30	405	541	1,519	2,034				
.10	.70	13	18	17	22	180	241	380	509	1,340	1,794		
.30	.70	13	18	17	22	85	114	173	231	593	793		
.50	.70	13	18	17	22	74	99	148	198	503	673		
.70	.70	13	18	17	22	85	114	173	231	593	793		
.90	.70	13	18	17	22	180	241	380	509	1,340	1,794		
.10	.80	10	14	13	17	102	136	169	226	335	449	1,090	1,459
.30	.80	10	14	13	17	48	64	77	103	149	199	475	635
.50	.80	10	14	13	17	42	56	66	88	126	169	401	536
.70	.80	10	14	13	17	48	64	77	103	149	199	475	635
.90	.80	10	14	13	17	102	136	169	226	335	449	1,090	1,459
.10	.90	8	11	10	13	65	87	95	128	149	200	273	365
.30	.90	8	11	10	13	31	41	44	58	66	89	119	159
.50	.90	8	11	10	13	27	36	37	50	56	75	101	134
.70	.90	8	11	10	13	31	41	44	58	66	89	119	159
.90	.90	8	11	10	13	65	87	95	128	149	200	273	365

[a] Calculations based on a goodness-of-fit formula provided by Donner and Eliasziw.[59]

When seeking to optimize sample size, the investigator needs to choose the appropriate balance between the number of raters examining each subject and the number of subjects.[60] In some instances, it is more practical to increase the number of raters rather than increase the number of subjects. However, according to Shoukri,[39] when seeking to detect a kappa of .40 or greater on a dichotomous variable, it is not advantageous to use more than 3 raters per subject—it can be shown that for a fixed number of observations, increasing the number of raters beyond 3 has little effect on the power of hypothesis tests or the width of confidence intervals. Therefore, increasing the number of subjects is the more effective strategy for maximizing power.

Conclusions

If used and interpreted appropriately, the kappa coefficient provides valuable information on the reliability of data obtained with diagnostic and other procedures used in musculoskeletal practice. We conclude with the following recommendations:

- Alongside the obtained value of kappa, report the bias and prevalence.
- Relate the magnitude of the kappa to the maximum attainable kappa for the contingency table concerned, as well as to 1; this provides an indication of the effect of imbalance in the marginal totals on the magnitude of kappa.
- Construct a confidence interval around the obtained value of kappa, to reflect sampling error.
- Test the significance of kappa against a value that represents a minimum acceptable level of agreement, rather than against zero, thereby testing whether its plausible values lie above an "acceptable" threshold.

- Use weighted kappa on scales that are ordinal in their original form, but avoid its use on interval/ratio scales collapsed into ordinal categories.
- Be cautious when comparing the magnitude of kappa across variables that have different prevalence or bias, or that are measured on different scales.

References

1 Toussaint R, Gawlik CS, Rehder U, Rüther W. Sacroiliac joint diagnostics in the Hamburg Construction Workers Study. *J Manipulative Physiol Ther.* 1999;22:139–143.

2 Fritz JM, George S. The use of a classification approach to identify subgroups of patients with acute low back pain. *Spine.* 2000;25: 106–114.

3 Riddle DL, Freburger JK. Evaluation of the presence of sacroiliac joint region dysfunction using a combination of tests: a multicenter intertester reliability study. *Phys Ther.* 2002;82:772–781.

4 Petersen T, Olsen S, Laslett M, Thorsen H, et al. Inter-tester reliability of a new diagnostic classification system for patients with non-specific low back pain. *Aust J Physiother.* 2004;50:85–91.

5 Fjellner A, Bexander C, Faleij R, Strender LE. Interexaminer reliability in physical examination of the cervical spine. *J Manipulative Physiol Ther.* 1999;22:511–516.

6 Hawk C, Phongphua C, Bleecker J, Swank L, et al. Preliminary study of the reliability of assessment procedures for indications for chiropractic adjustments of the lumbar spine. *J Manipulative Physiol Ther.* 1999;2:382–389.

7 Smedmark V, Wallin M, Arvidsson I. Inter-examiner reliability in assessing passive intervertebral motion of the cervical spine. *Man Ther.* 2000;5:97–101.

8 Hayes KW, Petersen CM. Reliability of assessing end-feel and pain and resistance sequences in subjects with painful shoulders and knees. *J Orthop Sports Phys Ther.* 2001;31:432–445.

9 Kilpikoski S, Airaksinen O, Kankaanpää M, et al. Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine.* 2002;27:E207–E214.

10 Hannes B, Karlmeinrad G, Ogon M, Martin K. Multisurgeon assessment of coronal pattern classifications systems for adolescent idiopathic scoliosis. *Spine.* 2002;27:762–767.

11 Speciale AC, Pietrobon R, Urban CW, et al. Observer variability in assessing lumbar spinal stenosis severity on magnetic resonance imaging and its relation to cross-sectional spinal canal area. *Spine.* 2002;27: 1082–1086.

12 Richards BS, Sucato DJ, Konigsberg DE, Ouellet JA. Comparison of reliability between the Lenke and King classification systems for adolescent idiopathic scoliosis using radiographs that were not premeasured. *Spine.* 2003;28:1148–1157.

13 Sim J, Wright CC. *Research in Health Care: Concepts, Designs and Methods.* Cheltenham, England: Nelson Thornes; 2000.

14 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.

15 Daly LE, Bourke GJ. *Interpretation and Uses of Medical Statistics.* 5th ed. Oxford, England: Blackwell Science; 2000.

16 Conger AJ. Integration and generalization of kappas for multiple raters. *Psychol Bull.* 1980;88:322–328.

17 Haley SM, Osberg JS. Kappa coefficient calculation using multiple ratings per subject: a special communication. *Phys Ther.* 1989;69: 970–974.

18 Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76:378–382.

19 Fritz JM, Wainner RS. Examining diagnostic tests: an evidence-based perspective. *Phys Ther.* 2001;81:1546–1564.

20 Feinstein AR, Cicchetti DV. High agreement but low kappa, I: the problems of two paradoxes. *J Clin Epidemiol.* 1990;43:543–549.

21 Bartko JJ, Carpenter WT. On the methods and theory of reliability. *J Nerv Ment Dis.* 1976;163:307–317.

22 Fleiss JL, Nee JCM, Landis JR. Large sample variance of kappa in the case of different sets of raters. *Psychol Bull.* 1979;86:974–977.

23 Hartmann DP. Considerations in the choice of interobserver reliability estimates. *J Appl Behav Anal.* 1977;10:103–116.

24 Rigby AS. Statistical methods in epidemiology, V: towards an understanding of the kappa coefficient. *Disabil Rehabil.* 2000;22: 339–344.

25 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull.* 1968;70:213–220.

26 Lantz C. Application and evaluation of the kappa statistic in the design and interpretation of chiropractic clinical research. *J Manipulative Physiol Ther.* 1997;20:521–528.

27 McKenzie RA. *The Lumbar Spine: Mechanical Diagnosis and Therapy.* Waikanae, New Zealand: Spinal Publications Ltd; 1981.

28 Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med.* 2002;21:2109–2129.

29 Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ.* 1992;304:1491–1494.

30 Donner A, Eliasziw M. Statistical implications of the choice between a dichotomous or continuous trait in studies of interobserver agreement. *Biometrics.* 1994;50:550–555.

31 Bartfay E, Donner A. The effect of collapsing multinomial data when assessing agreement. *Int J Epidemiol.* 2000;29:1070–1075.

32 Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol.* 1987;126:161–169.

33 Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to their Development and Use.* 3rd ed. Oxford, England: Oxford University Press; 2003.

34 Stratford PW, Goldsmith CH. Use of the standard error as a reliability index: an applied example using elbow flexor strength. *Phys Ther.* 1997;77:745–750.

35 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1(8476): 307–310.

36 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol.* 1993;46:423–429.

37 Bannerjee M, Fielding J. Interpreting kappa values for two-observer nursing diagnosis data. *Res Nurs Health.* 1997;20:465–470.

38 Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol.* 1988;41:949–958.

39 Shoukri MM. *Measures of Interobserver Agreement.* Boca Raton, Fla: Chapman & Hall/CRC; 2004.

40 Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. *Educ Psychol Meas.* 1981;41:687–699.

41 Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *J Clin Epidemiol.* 2000;53:499–503.

42 Cicchetti DV, Feinstein AR. High agreement but low kappa, II: resolving the paradoxes. *J Clin Epidemiol.* 1990;43:551–558.

43 Lantz C, Nebenzahl E. Behavior and interpretation of the κ statistic: resolution of the two paradoxes. *J Clin Epidemiol.* 1996;49:431–434.

44 Gjørup T. The kappa coefficient and the prevalence of a diagnosis. *Methods Inf Med.* 1988;27:184–186.

45 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.

46 Fleiss JL. *Statistical Methods for Rates and Proportions.* 2nd ed. New York, NY: John Wiley & Sons Inc; 1981.

47 Altman DG. *Practical Statistics for Medical Research.* London, England: Chapman & Hall/CRC; 1991.

48 Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res.* 1998;7:301–317.

49 Dunn G. *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors.* London, England: Edward Arnold; 1989.

50 Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology.* 1996;7:199–202.

51 Haas M. Statistical methodology for reliability studies. *J Manipulative Physiol Ther.* 1991;14:119–132.

52 Soeken KL, Prescott PA. Issues in the use of kappa to estimate reliability. *Med Care.* 1986;24:733–741.

53 Knight K, Hiller ML, Simpson DD, Broome KM. The validity of self-reported cocaine use in a criminal justice treatment sample. *Am J Drug Alcohol Abuse.* 1998;24:647–660.

54 Posner KL, Sampson PD, Caplan RA, et al. Measuring interrater reliability among multiple raters: an example of methods for nominal data. *Stat Med.* 1990;9:1103–1115.

55 Petersen IS. Using the kappa coefficient as a measure of reliability or reproducibility. *Chest.* 1998;114:946–947.

56 Main CJ, Wood PJ, Hollis S, et al. The distress and risk assessment method: a simple patient classification to identify distress and evaluate the risk of poor outcome. *Spine.* 1992;17:42–51.

57 Sim J, Reid N. Statistical inference by confidence intervals: issues of interpretation and utilization. *Phys Ther.* 1999;79:186–195.

58 Flack VF, Afifi AA, Lachenbruch PA, Schouten HJA. Sample size determinations for the two rater kappa statistic. *Psychometrika.* 1988;53: 321–325.

59 Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Stat Med.* 1992;11: 1511–1519.

60 Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med.* 1998;17:101–110.

Appendix.
Weighted Kappa

Formula for weighted kappa (κ_w):

$$\kappa_w = \frac{\sum wf_o - wf_c}{n - \sum wf_c}$$

where Σwf_o is the sum of the weighted observed frequencies in the cells of the contingency table, and Σwf_c is the sum of the weighted frequencies expected by chance in the cells of the contingency table.

Linear and quadratic weights are calculated as follows:

$$\text{linear weight} = 1 - \frac{|i-j|}{k-1}$$

$$\text{quadratic weight} = 1 - \left(\frac{i-j}{k-1}\right)^2$$

Where $i-j$ is the difference between the row category on the scale and the column category on the scale (the number of categories of disagreement), for the cell concerned, and k is the number of points on the scale.

The weights for unweighted, linear weighted, and quadratic weighted kappas for agreement on a 4-point ordinal scale are:

	Unweighted	Linear weights	Quadratic weights
No disagreement	1	1	1
Disagreement by 1 category	0	.67	.89
Disagreement by 2 categories	0	.33	.56
Disagreement by 3 categories	0	0	0

This method weighting is based on agreement; a method of weighing based on disagreement also can be used.[25]

A number of widely available computer programs will calculate κ and κ_w through standard options. SPSS 12[a] will calculate κ (but not κ_w) and performs a statistical test against a null value of zero. STATA 8[b] and SAS 8[c] calculate both κ and κ_w and perform a statistical test against a null value of zero for each of these statistics. PEPI 4[d] calculates both κ and κ_w, and provides a value of κ_{max} for each of these statistics. This program performs a statistical test against a null value of zero (and if the observed value of κ or κ_w >.40, against a null value of .40 also), together with 90%, 95%, and 99% 2-sided confidence intervals. Additionally, the prevalence-adjusted bias-adjusted kappa (PABAK) is calculated, and a McNemar test for bias is performed. Various other stand-alone programs are available, and macros can be used to perform additional functions related to κ for some of these programs.

Most programs will calculate a standard error of κ. Two standard errors can be calculated. A standard error assuming a zero value for κ should be used for hypothesis tests against a null hypothesis that states a zero value for κ, whereas a different standard error, assuming a nonzero value of κ, should be used for hypothesis tests against a null hypothesis that states a nonzero value for κ and to construct confidence intervals.

[a] SPSS Inc, 233 S Wacker Dr, Chicago, IL 60606.
[b] StataCorpLP, 4905 Lakeway Dr, College Station, TX 77845.
[c] SAS Institute Inc, 100 SAS Campus Dr, Cary, NC 27513.
[d] Sagebrush Press, 225 10th Ave, Salt Lake City, UT 84103.